SVEUČILIŠTE U ZAGREBU

FAKULTET ORGANIZACIJE I INFORMATIKE

V A R A Ž D I N

Vedran Grbavac

RUDARENJE PODATAKA KAO METODA UPRAVLJANJA ZNANJEM

ZAVRŠNI RAD

Varaždin, 2018.

SVEUČILIŠTE U ZAGREBU

FAKULTET ORGANIZACIJE I INFORMATIKE

V A R A Ž D I N

Vedran Grbavac

Matični broj: 44052/15-R

Studij: Informacijski sustavi

RUDARENJE PODATAKA KAO METODA UPRAVLJANJA ZNANJEM

ZAVRŠNI RAD

Mentor:

Izv. prof. dr. sc. Markus Schatten

Varaždin, rujan 2018.

Sadržaj

1. Uvod

U današnjem svijetu broj podataka rapidno raste te s time dolazi i do povećavanja njihove važnosti. Podaci se prikupljaju iz različitih izvora, kroz različite načine i u različitim oblicima. Također, njihovi izvori mogu biti javno dostupni ili mogu biti dostupni uz plaćanje nekakve naknade. Iz razloga što je broj podataka u svijetu se naglo povećava to stvara nove probleme za informacijske sustave koje je potrebno rješavati, ali s druge strane otvara i nove mogućnosti koje je potrebno otkriti. Kako bi uspješno rješavali probleme uz veliku količinu podataka razvile su se nove metode, alati i tehnike koje nam pomažu u tom procesu. Jedna od tih metoda je i rudarenje podataka koje obuhvaća istraživanje i analizu velikih količina podataka kako bi otkrili smislena pravila i uzorke, a konačnici rezultira dobivanjem novih znanja. Drugim riječima, rudarenjem nad već prikupljenim podacima dobijemo informacije od kojih možemo načiniti znanja.

Zbog puno podataka dobijemo puno informacija što u konačnici znači da možemo izvući puno znanja kojim je potrebno upravljati. Iz tog razloga razvila se disciplina pod nazivom upravljanje znanjem koji je danas veoma širok i teško objašnjiv pojam, ali se prakticira u mnogim organizacijama. Upravljanje znanjem ljudima daje mogućnost novog pogleda na nešto što su godinama gledali kao nebitne stvari. Također, ono nam daje odgovore na stvarne socijalne i ekonomske trendove, globalizaciju, informatizaciju, ali i centralistički pogled na znanje koje ima organizacija.

U ova dva prethodna odlomka napisano je i ukratko objašnjeno dva glavna pojma koja će se detaljno obrađivati tijekom ovog rada. Također, rad će se osim teoretskog dijela ovih pojmova još sastojati i od stvarnih primjera iz prakse te jednog moga primjera nad stvarnim podacima i donošenjem zaključaka na temelju provedene analize.

2. Upravljanje znanjem

Upravljanje znanjem ima veoma veliku i daleku povijest, sve je počelo sa razgovorom između ljudi na radnom mjestu pa se proširila sve do profesionalnog treninga te programa s mentorom, a njezin razvitak se očekuje još dugi niz godina u budućnosti. Kada govorimo o upravljanju znanja postoji niz različitih definicija koje opisuju taj pojam. Kratka definicija za to bi bila da je to niz postupaka koje koriste organizacije kako bi prepoznale, stvorile, prezentirale i distribuirale znanje u ponovne svrhe, ovu definiciju ćemo proširiti u nastavku ovog poglavlja kada najprije objasnimo neke od pojmova koji su nam potrebni kako bi lakše shvatili i razumjeli upravljanje znanjem.

2.1. Podatak, informacija i znanje

Upravljanje znanjem nema nikakvog značaja ako ne napravimo razliku između podatka, informacije i znanja. U organizacijama većinom je menadžerski posao da raspoznaju te tri različite stvari te da na principu njih donose kvalitetne odluke o njihovom korištenju kako bi bile korisne organizaciji.

2.1.1. Podatak

Podatak je činjenica predočena u formaliziranom obliku koja je pogodna za komunikaciju, interpretaciju i obradu uz pomoć ljudi ili strojeva. U osnovi podatak je nekakva poruka koja može, ali ne mora koristiti, npr. kao broj, slika ili riječ. Primjer za podatak u stvarnom svijetu za trgovca bio bi kupci, prodajna mjesta ili roba.

2.1.2. Informacija

Informacija je značenje koje čovjek ili stroj pripisuje podatku, odnosno, to je rezultat obrade, manipulacije i organizacije podataka na način koji dodaje znanje primatelju. Ukratko, kada podatak iskoristimo u neku korisnu svrhu onda je to informacija. U stvarnom svijetu primjer za informaciju bi bio izvješće prodaje za prethodni mjesec, u biti to je papir sa listom brojeva, ali za voditelja prodaje to predstavlja značaju informaciju koju može izvaditi iz sirovih podataka.

2.1.3. Znanje

Pojam znanje koristimo za ispravnu primjenu informacija, odnosno, to je personalizirana informacija koju posjeduje pojedinac, a povezana je sa njegovim kognitivnim sposobnostima te sposobnostima analize i procjene. Sa gledišta organizacije znanje predstavlja iskoristive intelektualne resurse. Ukratko, informacija se pretvara u znanje kada se koristi za donošenje odluka te planiranje odgovarajućih akcija. Razlika između informacije i znanja objašnjena je u nastavku kroz primjer plesa. Informacija za plesača odnosi se na korake koje mora napraviti, dok se znanje odnosi na to kako će plesači izvesti te korake, odnosno, kako će usavršiti cijeli ples. Još jedan primjer iz stvarnog svijet bi bila analiza prodaje poduzeća koja nam može pokazati npr. na kojim područjima se najmanje robe prodaje, koja dob ljudi najviše kupuje našu robu i slično te na principu tih podataka zaduženi ljudi u organizaciji mogu donijeti odluke koje će popraviti tu statistiku kao što su npr. uvesti besplatnu dostavu za područja koja najmanje kupuju te u konačnici osigurati da poduzeće poveća svoju prodaju i u budućnosti raste u svakom segmentu poslovanja.

U praktičnoj primjeni znanja moramo biti sigurni u vlastitu mogućnost raspoznavanja podataka i informacija od primjenjivog znanja. U organizacijama baze podataka mogu biti pune informacija, ali one nemaju nikakve koristi od njih ukoliko ne postoji osoba koja na temelju njih može donijeti korisne i dobre odluke za poduzeće. Drugim riječima, informacija postaje znanje tek kada postoji osoba koja zna protumačiti i primijeniti tu informaciju. U stvarnom svijetu moderne organizacije imaju pristup raznim repozitorijima koji su puni informacija, ali proces otkrivanja znanja u podacima i dalje ima probleme, neki od njih su:

- Manjak informacija – događa se kada tražimo informacije za određenu temu za koju znamo da postoje informacije, ali ih mi ne možemo pronaći
- Previše informacija - otkrivanje novog znanja nam oduzima previše vremena ako pronađemo previše informacije o određenom objektu koje nemaju nikakvu vrijednost

Slika 1: Vrijednost znanja [Fernandez, Gonzalez, i Sabherwal, 2004, str.15]

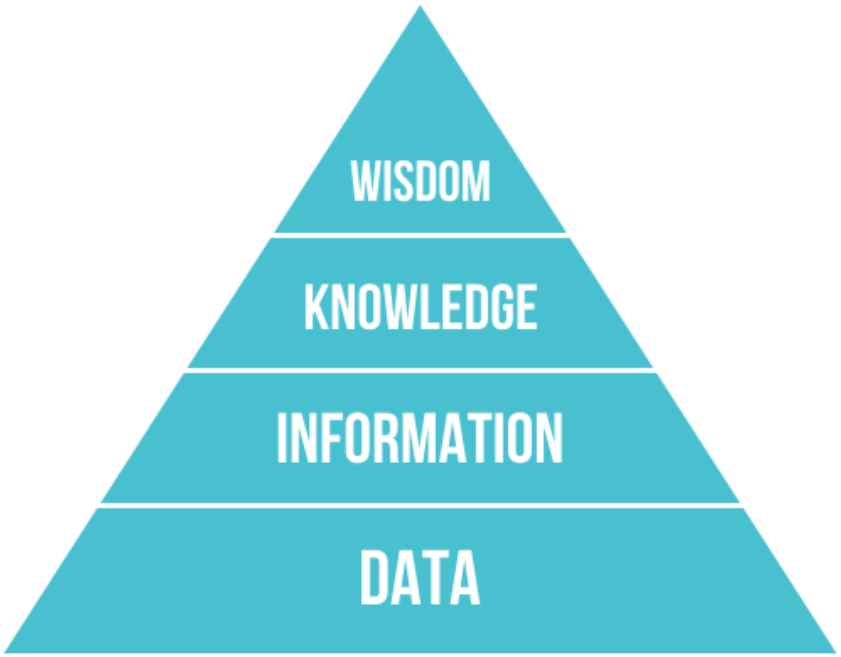
Razlikujemo različite tipove znanja kao što su: opće, specifično, proceduralno, deklarativno, tacitno, eksplicitno te ostale tipove znanja kao što su: jednostavno, složeno, strateško, heurističko, strukturna, neegzaktna, ustaljena itd.

Opće ili kako ga još zovemo generalno znanje obuhvaća znanja koja su lako prenosiva i posjeduje ih veliki broj ljudi, primjer za to znanje je da veliki broj ljudi zna da je u hladu hladnije. S druge strane imamo specifično znanje koje posjeduju razni stručnjaci tj. posjeduje ga veoma mali broj ljudi te je teško prenosivo. Primjer za specifično znanje bilo bi kako uz pomoć *„big data"* utjecati na ljude. Proceduralno znanje je znanje o nekom postupku, odnosno, kako izvršiti neki zadatak, točnije rečeno koje korake treba izvršiti i kojim redom, te koja pravila poštivati, ne podrazumijeva nužno razloge zašto to činimo te kako to utječe na okolinu. Primjer za proceduralno znanje bi bilo povezivanje pokreta u plivanju tj. kojim redom i kako moramo obavljati određene korake kako bi ostali na površini vode i kretali se prema naprijed. Deklarativno znanje opisuje ono što je poznato u vezi s problemom. Stručnije rečeno, to je znanje o vezama između varijabli. Laički rečeno, razumijevanje zadatka koje radimo, poznatim objektima, konceptima i činjenicama, znamo točno što radimo, zašto radimo te kako ti postupci utječu na radnu okolinu. Primjer za to bilo bi znanje o plivačkim pokretima. Tacitno znanje je utemeljeno na individualno iskustvu, intuiciji, razumijevanju i predosjećaju osobe te ga je teško objasniti i prenijeti. Ovo znanje puno se lakše prenosi pokazivanje kako nešto napraviti nego objašnjavanje riječima ili čitanjem iz knjiga. Za ovo znanje još se kaže da je skriveno znanje, odnosno, da nije dokumentirano i svima dostupno, a iskazuje se kroz vještinu, intuiciju i iskustvo individue. Primjer za ovo bi bio vožnja bicikla jer kao što znamo nitko ne čita knjige o tome kako se treba voziti bicikli nego nam neko stariji i iskusniji pokaže kako se to radi te nakon nekog vremena i mi samo to svladamo. Eksplicitna znanja su znanja koja lako naučimo, može se prenijeti riječima i brojevima ili naučiti iz knjige te su dostupna svima. Primjer za ovo znanje su upute za korištenje u kojima sve piše što, kako i zašto nešto koristiti, a drugi primjer za ovo bi bio i udžbenici koje se koriste u školama ili fakultetima. [Woods i Cortada, 2000., str. 14-16 , prezentacije kolegija *„Upravljanje znanjem"*]

2.1.4. Podatak, informacija, znanje, mudrost(eng. DIKW hierarchy)

Ovaj pojam također je poznat i pod nazivom DIKW pyramid, odnosno na hrvatskom DIKW piramida. Slova DIKW zapravo predstavljaju podatak(eng. *data*), informaciju(eng. *information*), znanje(eng. *knowledge*) i mudrost(eng. *wisdom*) te nam ova piramida zapravo predstavlja odnos između toga. Postoji više vrsta ovakve piramide te se sve one ne obuhvaćaju sve četiri stavke nego samo neke od njih dok mogu i dodavati neke nove. Obično

se informacije definiraju u smislu podatka, znanja u smislu informacija te mudrosti u smislu znanja. Za što nam ovo točno koristi? DIKW je prijedlog strukturiranja podataka, informacija, znanja i mudrosti u jednu informacijsku hijerarhiju gdje svaka razina dodaje određena svojstva iznad i ispod one prethodne. Krećemo od podatka koji je najosnovnija razina te idemo prema informaciji koja dodaje nekakav kontekst zatim slijedi znanje koje govori kako ga upotrijebiti i na kraju mudrost koja dodaje kada i zašto ga upotrijebiti.



Slika 2: DIKW piramida

2.2. Definicija upravljanja znanjem

Znanje se ne može upravljati, ono postoji unutar uma pojedinca. (Groff i Jones, 2003., str. 2.). Iz tog proizlazi da bi definicija upravljanja znanja bi morala uključivati komponentu znanja i komponentu vlasnika tog znanja. Kao što je već spomenuto da postoji puno definicija koje objašnjavaju što je zapravo upravljanje znanje te ako ćemo iskoristi neke od njih te uz pomoć njih pokušati što bolje objasniti što je upravljanje znanjem zapravo.

> *„Upravljanje znanjem je niz međusobno povezanih aktivnosti organizacija i menadžmenta usmjerenih na strategiju i taktiku upravljanja ljudskim kapitalom, odnosno razvoj znanja, vještina i općenito kompetencija zaposlenih, KNOW-HOW-a, kroz obrazovanje i obuku, stjecanje radnog i profesionalnog iskustva i slično.“ (Bahtijarević – Šiber i Sikavica, 2001., str. 629. – 630.)*

Autori ove definicije su puno toga rekli u jedno rečenici te ćemo u nastavku ovog teksta to sve objasniti dio po dio. Oni govore da upravljanje znanjem nije jednostavna aktivnost nego veoma komplicirana koja se sastoji od mnogo drugi aktivnosti koje su međusobno povezane te su usmjerene na razvoj organizacije. Prvenstveno je usmjereno na razvoj znanja i kompetencija svojih zaposlenika te da oni znaju kako(KNOW-HOW) učiniti nešto na najbolji i najprihvatljiviji način. Također, govore o tome kako se znanje treba povećavati svakodnevno, odnosno, dobiti početno znanje kroz obrazovanje i obuku, a onda kroz radno iskustvo to znanje proširivati i stjecati mudrost.

> *“Upravljanje znanjem je sustavni postupak uspostave, održavanja i usmjeravanja cjelokupne organizacije u svrhu korištenja znanja radi stvaranja poslovnih vrijednosti i generiranja konkurentske prednosti.” [Žugaj i Schatten, 2005., str. 67]*

Prvi korak u organizaciji bi uvijek trebao biti njeno održavanje te strategijsko planiranje opstanka. Zvuči veoma jednostavno, no međutim to nije lagan zadatak za organizaciju te zbog toga strategija razvoja organizacije mora biti temeljita i razumljiva svima. U oblikovanju strategije znanje je od velikog značaja te kroz vrijeme postaje sve veće i veće te upravo zbog toga upravljanje znanjem rezultira stvaranjem poslovnih vrijednosti organizacije te njezinom konkurentnošću na tržištu. Suvremene organizacije svakodnevno dobivaju velike količine podataka s kojima je potrebno upravljati da bi mogli iz njih izvući nekakvu korist. Veliki broj ti podataka je zapravo beskoristan za organizaciju, ali svejedno ih ne treba zanemariti jer i na

temelju njih organizacija može ipak nešto izvući što bi joj moglo biti od koristi. Na primjer to mogu biti nekakva prošla iskustva zbog kojih je sustav danas prilagođen jer je u prošlosti zakazao na tom dijelu te si ne želimo dopustiti ponovno iste gubitke. Zbog te ogromne količine podataka organizacije koriste raznorazne informatičke sustave koji im uvelike olakšavaju posao. Uz pomoć sustava i stručnjaka za određena područja podaci i informacije se čitaju, obrađuju, spajaju i slično te upravo to predstavlja manipulacije, odnosno, upravljanje podacima kako bi došli do nekakve koristi od tih informacija i tek to smatramo znanjem.

> *„Upravljanje znanjem jednostavno se može definirati kao poduzimanje potrebnih postupaka da se ostvari maksimalna korist iz resursa znanja.“ [Fernandez, Gonzalez, i Sabherwal, 2004., str.2]*

U organizacijama upravljanjem znanjem primarno se koriste menadžeri i upravitelji. Menadžment je proces kojim utječemo i upravljamo okruženjem i ljudima da bi lakše ostvarili zadane ciljeve. Kao što možemo pretpostaviti glavni cilj menadžmenta je povećanje profitabilnosti organizacije, a to postižu povećavanjem efektivnosti i efikasnosti organizacije stvarajući nove resurse te uz pomoć njih otvaraju nove mogućnosti. Menadžeri moraju donositi točne, sigurne i pravodobne odluke kako bi povećali efektivnost te moraju znati prepoznati važnost postojećih poslova te da ne troše resurse na ostvarivanje manje bitnih ciljeva. Oni daju točne smjernice kako određene poslove obaviti na što brži i bolji način te postavljaju ljude s najboljim kompetencijama na njima odgovarajuće poslove i omogućuju im pristup potrebnim resursima i informacijama. Upravo u ovom dijelu menadžeri se koriste upravljanjem znanjem u svrhu promicanja vlastitog poslovanja što zapravo i je jedna od glavnih svrha upravljanja znanjem jer ono osigurava da bitno znanje bude raspoloživo kad god i gdje kod je potrebno.

Još jednom za kraj da se naglasi da postoji puno definicija koje opisuju upravljanje znanjem te se svaka od njih razlikuje i ima svoje prednosti. U budućnosti sigurno će razni autori smisliti još neke definicije, ali ipak svi autori su složni kod definiranja cilja upravljanja znanjem. Cilj upravljanja znanja je osigurati eksplicitno i tacitno znanje te ostvariti uvjete za inovacije u svrhu kvalitetnijeg procesa donošenja odluka. (Wiig, 2004., str. 78.; Snowden, 2003., str. 113.)

3. Rudarenje podataka

U današnje vrijeme velike organizacije primaju te skupljaju veliku količinu podataka iz kojih je teško izdvojiti ono bitno te na temelju toga donositi kvalitetne poslovne odluke. Iz navedenog razloga poduzeća su počela koristiti rudarenje podataka kako bi si olakšali posao te iz tog velikog skupa podataka uz pomoć raznih metoda došli do potrebnih informacije, odnosno, znanja. Ukratko rečeno, rudarenje podataka je sortiranje, organiziranje ili grupiranje velikog broja podataka i izvlačenje relevantnih informacija.

3.1. Definicija rudarenja podataka

Rudarenje podataka još poznato kao i otkrivanje znanja u bazama podataka(eng. Knowleadge discovery in databases – KDD) je proces otkrivanja zanimljivih uzoraka većinom na velikom skupu podataka koji mogu biti od koristi organizaciji, točnije, uz pomoć koji organizacija može donijeti kvalitetne poslovne odluke za rast i razvoj organizacije. Pojam rudarenja podataka možemo razumjeti kroz širi ili uži način. Šire shvaćanje obuhvaća cjelokupni proces otkrivanja znanja iz dostupnih podataka, dok uži način obuhvaća specifičnu fazu obrade podataka. Podaci koji se obrađuju mogu biti tekstualni podaci, nestrukturirani podaci, podaci organizirani u vremenske sesije ili podaci organizirani u baze podataka koji je ujedno i najčešći oblik. Uz pojam rudarenje podataka postoje još i nazivi ekstrakcija znanja, analiza obrazaca, žetva informacija te cijeđenje podataka. (Srića, 2018.; Garača i Jadrić, 2011.)

> *„Rudarenje podataka je istraživanje i analiza velikih količina podataka u nastojanju otkrivanja smislenih obrazaca i pravila.“ (Berry i Linoff, 2004.).*

Prema ovome možemo zaključiti da je rudarenje podataka najčešće vezano uz velike količine podataka te njihovo istraživanje i analiziranje veza, ali i zakonitosti u podacima koristeći različite metode i tehnike kako bi iz strukturiranih ili nestrukturiranih skupova podataka dobili jasan pogled na njihovu međusobnu povezanost, odnosno, kako bi mogli otkriti neko znanje iz tih podataka. Uzeći to u obzir, može se reći da rudarenje podataka nema smisla na malim količinama podataka jer većina algoritama za rudarenje podataka zahtjeva velike količine podataka kako bi se mogli izgraditi modeli koji će koristiti za klasificiranje, procjene ili neke druge zadatke rudarenja podataka. (Garača i Jadrić, 2011.)

„Rudarenje podataka je pribavljanje ili „rudarenje“ znanja iz velikih izvora podataka.“ (Han i Kamber, 2006.)

Danas, rudarenje podataka postaje sve korisnije te velike organizacije teško mogu donijeti odluke bez toga. Također, ono se ne koristi samo za dobivanje poveznica između skupova podataka nego se koristi za poboljšavanje velikog broja drugih usluga te tako u komercijalnim sustavima služi da poboljšanje i povećavanje prodaje robe ili za bolje donošenje određenih teorija u istraživačkim centrima. Osim toga, koristi se i u vojsci za otkrivanje neprijateljskih baza, logora, povezivanje više različitih događaja, otkrivanja krivaca za prošle napade te na temelju toga donositi zaključke o mogućim budućim napadima.

Ovako rudarenje podataka kakvo danas postoji ne bi bilo moguće bez novih tehnologija i algoritama u informatici kao što su npr. multiprocesorska računala, moćni današnji serveri, niz tehnologija za masivno prikupljanje podataka te algoritamske tehnike koje omogućuju rudarenjem ogromnom količinom podataka.

Možemo reći da je rudarenje podataka izrazito multidisciplinarno područje koje može obuhvaćati puno različitih područja kao što su ekonomija, medicina, genetika, mikrobiologija, mehanika, farmacija te s druge strana još obuhvaća i područja matematike, statistike, baze podataka, teoriju informacija i slična pridružena područja.

Na konferencijama kojima je glavna tema poslovna inteligencija i upravljanje znanjem predstavljaju se projekti vodećih hrvatskih poduzeća kojima je rudarenje podataka sastavni dio organizacije te zbog toga sa sigurnošću možemo reći da je rudarenje podataka i u Hrvatskoj doživjelo pravi procvat. (Pejić-Bach, 2005.)

Kao što je već navedeno i ranije u tekstu za rudarenje podataka postoji čitavi niz faktora koji mogu utjecati na konačni rezultat događaja, ali ipak glavni zadatak rudarenja podataka je otkriti one najznačajnije faktore te njihove karakteristike s obzirom na ciljna stanja. (Klepac, 2006.)

Ono što je nekada bilo upravljanje znanjem, točnije inženjerstvo znanja danas je to rudarenje podataka. Do prije nekoliko godina ovo je bila nejasna i egzotična teorija o kojoj su raspravljali samo teoretičari dok je to danas sastavni dio svake veće i jače organizacije. Često možemo čuti da se rudarenje podataka poistovjećuje sa procesima otkrivanja i predviđanja znanja. Na kraju, možemo sa sigurnošću reći da rudarenje podataka privlači sve više i više pažnje u poslovnom svijetu što se može vidjeti i prema broju novih alata na tržištu koji

svakodnevno rastu ili člancima u popularnim IT časopisima. Također, možemo vidjeti da veliki broj organizacija počinje shvaćati da ogromne količine podataka o njihovim klijentima i njihovim ponašanjima sadrže vrijedne informacije za daljnji rast iste organizacije.

3.2. Zadaće rudarenja podataka

Postoji šest zadaća rudarenja podataka koje ćemo ovdje samo navesti, a u nastavku ovog poglavlja ćemo te iste zadaće detaljno objasniti, one su:

- Deskripcija
- Klasifikacija
- Estimacija(regresija)
- Predikcija
- Klasteriranje
- Asocijacija

3.2.1. Deskripcija

Istraživači i analitičari u rijetkim slučajevima ne mogu pronaći načine kako bi opisali neke pravilnosti i trendove koje su izvukli iz podataka. Pojasniti ćemo to bolje na primjeru, recimo da provodimo istraživanje nad ljudima koji bi na izborima ponovno dali glas za vladajuću stranku. Sada uzmimo u obzir skupinu ljudi koji su dobili otkaz, oni su trenutno u lošijem financijskom stanju zbog nedostatka posla pa je logično zaključiti da bi oni vjerojatno dali glas nekoj drugoj stranci jer žele promjenu. Opisi trendova i pravilnosti često nam daju moguća objašnjenja za takve pravilnosti i trendove te zbog toga KDD modeli moraju biti što je više moguće transparentni. Laički rečeno, rezultati naših modela moraju opisivati jasne pravilnosti koje svatko može intuitivno interpretirati i objasniti. Kod rudarenja podataka postoji više vrsta metoda o kojima ću detaljno pisati u nastavku ovog rada te neke od tih metoda su prikladnije od drugih što se tiče transparentne interpretacije. Na primjer, stabla odlučivanja daju intuitivno rješenje koje ljudi mogu razumjeti dok neuronske mreže su većinom crna kutija koju razumiju samo stručnjaci u tom području zbog njezine kompleksnosti.

3.2.2. Klasifikacija

Vrijednost koja se predviđa nazivamo ciljna varijabla te kod klasifikacije ona predstavlja određenu kategoriju. Uzmimo za primjer mjesečni prihod naših stanovnika, on se može podijeliti u tri kategorije: niski, srednji i visoki. Pošto rudarenje podataka radi sa velikom

količinom podataka, naš model čita veliki broj zapisa pri čemu svaki taj zapis sadrži informaciju o ciljnoj varijabli te skup ulaznih, odnosno, prediktorskih varijabli. Primjer prediktorskih varijabli su dob, spol ili zanimanje. Na temelju njihovih vrijednosti za već postojeće zapise u našoj bazi, analitičar pokušava uz pomoć modela odrediti koliki bi bio prihod za osobe koje nemamo zapisane u bazi.

Općenito algoritam za klasifikaciju bi išao ovim redoslijedom:

1. Pregledava skup podataka koji sadrži ciljnu varijablu te prediktorske varijable
2. Uči koje prediktorske varijable su povezane sa kojim ciljnim varijablama
3. Pregledava zapise koji nemaju zabilježene vrijednosti te na temelju prethodnog učenja pridjeljuje im određene klasifikacije

Podaci na kojima se naš algoritam uči još se nazivaju i podaci za trening. Primjeri gdje se koristi klasifikacija su: određivanje je li neka transakcija s kreditnom karticom legalna, određivanje je li oporuka napisana od strane pokojnika ili je krivotvorena te kod dijagnosticiranja određene bolesti kod pacijenata.

3.2.3. Estimacija(regresija)

Ova zadaća je vrlo slična prethodnoj s jednom bitnom razlikom, a to je da ciljna varijabla nije kategorijska nego je numerička. Pomoću kompletnih zapisa tj. onih zapisa koji sadržavaju ciljne varijable uz prediktorske varijable izrađujemo modele te se nakon toga za nove opservacije procjenjuje vrijednost ciljne varijable na temelju vrijednosti prediktorski varijabli. Većina regresijskih modela dolazi iz područja statičke analize. Neki od primjera za regresiju su: procjena postotka smanjenja brzine trkača nakon ozljede koljena ili procjena prosječne ocjene srednjoškolaca na temelju ocjena iz osnovne škole.

3.2.4. Predikcija

Vrlo je slična klasifikaciji i estimaciji s time da u ovoj zadaći ciljna varijabla predstavlja buduću varijablu što i samo ime govori, predikcija. Metode koje smo koristili za klasifikaciju i estimaciju također se mogu koristiti i za predikciju pod pravim okolnostima. U te metode možemo uključiti i tradicionalne metode kao što su jednostavna linearna regresija i korelacija, višestruka regresija, ali se također mogu i uključiti metode rudarenja podataka kao što su neuronske mreže ili stabla odlučivanja. Primjer za predikciju je predviđanje cijene dionica sljedeća tri mjeseca.

3.2.5. Klasteriranje

Kao što joj i samo ime kaže, ova zadaća se odnosi na grupiranje zapisa, opservacija ili slučajeva u klastere sličnih objekata. Klaster je skup zapisa koju su međusobno slični, ali se razlikuju po nečemu od zapisa u drugim klasterima. Za klasteriranje ne postoji ciljna varijabla te se ono ne koristi za klasificiranje, procjenu ili predviđanje ciljne varijable. Algoritam za ovu zadaću radi na način da pokušava segmentirati cijeli skup podataka u relativno homogene podgrupe, odnosno, klastere s tim da unutar klastera sličnost zapisa bude maksimalna dok sličnost za zapisima izvan tog klastera bude minimalna. Ovaj način većinom se koristi kao nekakav početni korak u KDD procesu, a rezultirajući klasteri se koriste kao dodatni ulazni atributi za daljnje tehnike rudarenja podataka. Primjer za ovu zadaću bi bio: smanjenje broja dimenzija skupa podataka sa stotinama dimenzija(atributa).

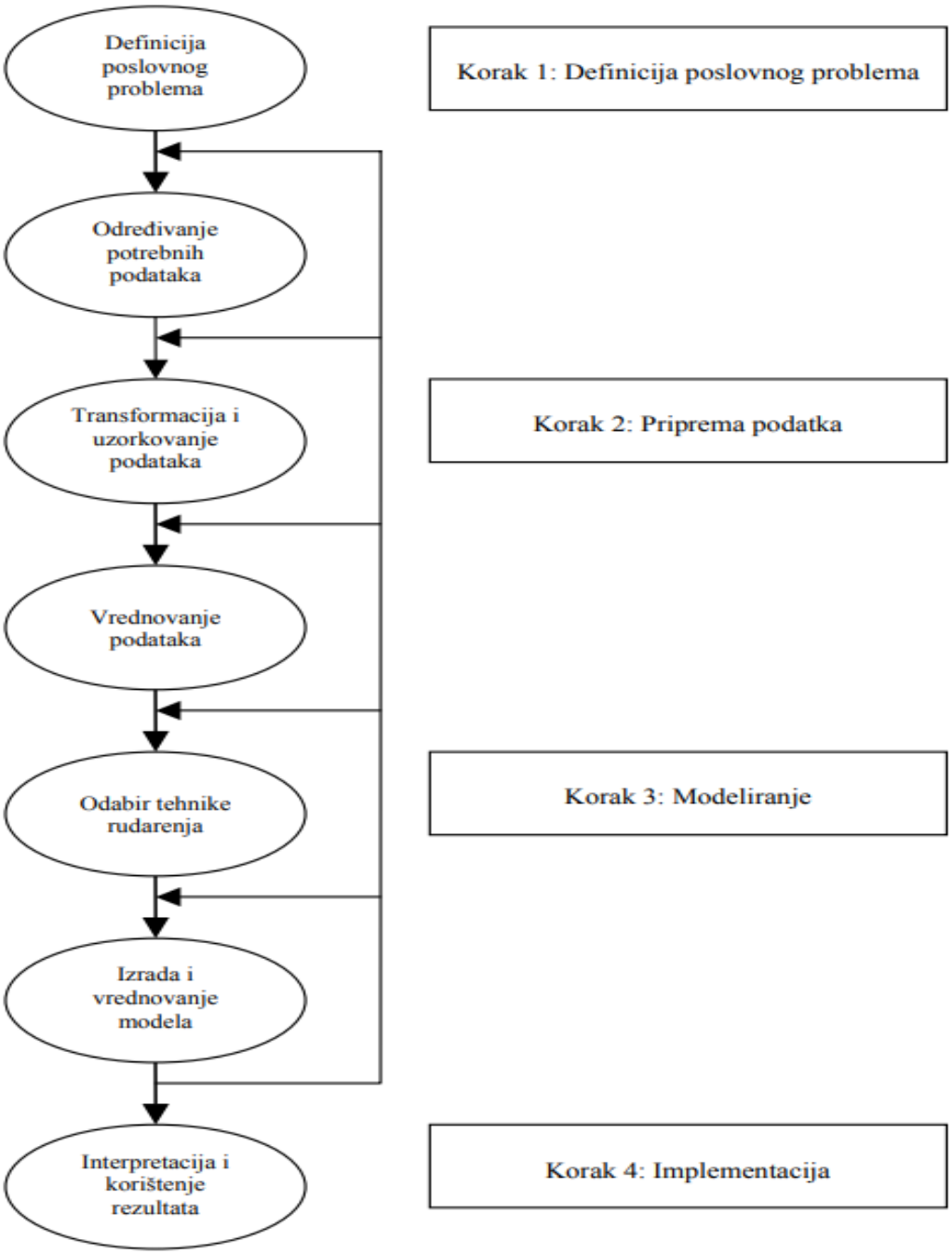
3.2.6. Asocijacija

Cilj ove zadaće je otkriti koji su atributi međusobno povezani. Ova zadaća se često koristi u poslovnom svijetu gdje se još naziva analiza afiniteta ili analiza potrošačke košare. Pravila za asocijaciju su oblika *„ako uvjet“* onda *„posljedica“* te također ima mjeru značaja i pouzdanosti. Objasnimo to na jednom primjeru, recimo da trgovina provodi istraživanje nad svojim kupcima te su uočili da petkom navečer od 400 kupaca, 100 ih je kupilo alkohol te od tih 100, 60 ih je kupilo cigarete. Dakle izračun i asocijacijsko pravilo bi izgledalo ovako *„Ako je kupac kupio alkohol, kupit će i cigarete“* sa značajem od 100/400 = 25% i pouzdanosti od 60/100 = 60%. Još neki primjeri koji uključuju asocijaciju: istraživanje proporcije djece čiji roditelji im čitaju da i oni sami čitaju ili koji artikli se u trgovini prodaju zajedno, odnosno, oni koji se nikad ne prodaju zajedno.

3.3. Faze rudarenja podataka

Važno je istaknuti da za uspješno rudarenje podataka koje će nam donijeti puno vrijednih informacija ne postoji unikatan način, ali vjerojatnost njena uspjeha možemo povećati sljedeći neke od koraka procesa rudarenja podataka.

Prema Pejiću i Bachu cjelokupan proces rudarenja podataka bi trebao uključivati četiri faze, koje su prikazane na slici 2.



Slika 3: Proces rudarenja podataka

3.3.1. Definicija poslovnog problema

Prvi korak kao što i samo ime kaže nam definira naš poslovni problem koji je potrebno izraziti u obliku pitanja na koja ćemo moći odgovoriti po završetku procesa. Da bi ovaj korak izveli najbolje što možemo potrebno je prvo provesti analizu područja gdje je rudarenje podataka već uspješno korišteno. Također, potrebno je odrediti ljude koji će raditi na ovom problemu, obično to bude manja skupina ljudi koja uključuje specijalista za rudarenje podataka koji dobro poznaje metode otkrivanja znanja, informatičar koji ima iskustva sa radom sa bazom i skladištima podataka te stručnjak iz organizacije koji je dobro upoznat sa potencijalnim primjerom u poslovanju. Osim navedenih ljudi, potreban nam je menadžer koji će voditi ovaj tim na način da ne mora direktno raditi na njemu, ali treba im pomoći u rješavanju eventualnih problema.

3.3.2. Priprema podataka

Druga faza obuhvaća pripremu podataka koja se i sama razlaže na nekoliko koraka. Priprema podataka uključuje određivanje potrebnih podataka, transformaciju, uzrokovanje i vrednovanje podataka. Ova faza je vremenski najduža, točnije prema nekim autorima ona oduzima od 60% do 90% ukupnog vremena rudarenja podataka. Najčešće podaci su smješteni unutar baza ili skladišta podataka, ali ponekad mogu biti pohranjeni i u nekim drugim oblicima. Ljudi koji rade na projektu rudarenja podataka moraju zajedno odrediti koji će podaci biti potrebni za izradu modela, a koje podatke treba izbaciti.

3.3.2.1. Određivanje potrebnih podataka

Ovo je jedan od koraka pripreme podataka, a u njemu određujemo koje ćemo varijable uzeti za ciljne, odnosno, zavise, a koje varijable treba izbaciti. Na primjeru analize kreditnog rizika, naša ciljna varijabla biti će on koja nam govori je li klijent vratio kredit ili nije. Na kraju ovog koraka ćemo dobiti popis varijabli koje ćemo dalje koristiti za izradu modela.

3.3.2.2. Transformacija podataka

Sljedeći korak kao što mu i samo ime kaže odnosi se na pretvorbu podataka u oblik koji će nam biti pogodan za rudarenje. Podaci za rudarenje moraju biti u tabličnom obliku tako da stupci predstavljaju varijable dok su u recima opažanja. Također, ovaj korak obuhvaća i operacije s podacima kao što su agregacija, grupiranje, selekcija, filtriranje i spajanje.

3.3.2.3. Uzrokovanje podataka

Ovaj korak bi nam u konačnici trebao odgovoriti na pitanje *„Koliko podataka je dovoljno?“*. Jedinstveni odgovor na ovo pitanje ne postoji jer količina podataka ovisi od algoritma do algoritma. Kao što smo već upoznati da baze podataka sadrže ogromne količine podataka, zadatak uzrokovanja je da pokuša ta količina što više smanjiti za daljnju izradu modela. U stvarnom svijetu, za metodu stabla odlučivanja potrebno je od dvije do tri tisuće podataka dok za metodu neuronskih mreža ta brojka raste preko deset tisuća. Odabir podataka za uzorak najčešće se izabire slučajnim odabirom te nakon što je izabran uzorak još ga je potrebno podijeliti na dio podataka za izradu modela i dio podataka za testiranje modela. Ovo je tipičan pristup rudarenja podataka jer na ovaj način provjeravamo njegovu efikasnost na podacima koji nisu izabrani za njegovu izradu.

3.3.2.4. Vrednovanje podataka

Zadnji korak pripreme podataka kojemu je cilj da izbaci sve one „prljave“ podatke kako bi dobili „čisti“ model. Prljave podatke nazivamo one podatke koji su netipične vrijednosti, odnosno, nepostojeće, netočne ili nejasne vrijednosti, a javljaju se u našim bazama podataka. Oni najčešće nastaju prelaskom iz jedne baze podataka u drugu ili su posljedica pogrešnog unosa podataka u računalo.

3.3.3. Modeliranje

U ovoj fazi moramo odabrati metodu rudarenja podataka te izraditi i vrednovati model. U procesu rudarenja podataka koristimo nekoliko metoda:

- Statistika
- Umjetna inteligencija
- Baze i skladišta podataka
- Predviđanje

Metode rudarenja podataka dijelimo u tri kategorije:

1. Otkrivanje
2. Kvalifikacija
3. Predviđanje

Nakon što primijenimo metode, njihove rezultate vrednujemo uz pomoć podataka za testiranje modela, više o modeliranju u sljedećem poglavlju.

3.3.4. Implementacija i korištenje rezultata

Zadnji korak kao što ste mogli pretpostaviti odnosi se na implementaciju i korištenje rezultata dobivenih kroz modele. Od velike je važnosti da rezultati modela budu u jednostavnom i razumljivom obliku za interpretaciju kao što su na primjer grafikoni ili pravila. Razlog zašto moraju baš takvi biti je taj što korisnik koji čita te rezultate nije nikakav stručnjak nego obična osoba koja se služi našim rezultatima. Također, bitno je da rezultati budu što bolje predstavljeni jer će se na taj način sve više koristiti.

Proces rudarenja podataka je iterativan proces, odnosno, sastoji se od puno ponavljanja te je zbog toga uvijek moguće se vratiti nekoliko koraka unatrag ako nešto nije u redu. Recimo, da smo u trećem koraku shvatili da podaci koje koristimo nisu dobro odabrani onda ćemo se vratiti jedno korak unatrag, odabrati odgovarajuće podatke te ponovno se vratiti na izradu modela. Greške u koracima su česta stvar te ih se ne treba bojati, nego jednostavno što se više vraćamo to će naši podaci biti puno bolji te samim time model i njegovi rezultati puno kvalitetniji.

3.4. Metode rudarenja podataka

Polovicom devedesetih godina prošlog stoljeća dolazi do uređenja u području rudarenja podacima koje objedinjava skup metoda i postupaka koji za cilj imaju otkriti određene zakonitosti u ogromnim količinama podataka, ali tek u zadnjih desetak godina komercijalno rudarenje dobiva strateški značaj u poslovni organizacijama. (Garača i Jadrić, 2011.; Klepac, 2006.)

Već je u ovom radu spomenuto, ali naglasiti ćemo ponovno da metode rudarenja podataka svoje korijene vuku iz drugih područja kao što su statistika, teorija informacija, baza podataka, teorija vjerojatnosti i umjetna inteligencija. Ne postoji univerzalna metoda koja će nam uvijek davati kvalitetne rezultate, nego ovisno o prirodi problema, dostupnosti podataka i sklonostima

izvođača mi odabiremo neku od metoda koja nama najviše odgovara u tom trenutku. Veliki dio rudarenja podataka oslanja se na izradu kvalitetnog modela, a taj naš model zapravo predstavlja algoritam, odnosno, skup pravila koja povezuju ulaze s određenom ciljnom varijablom. (Garača i Jadrić, 2011.; Klepac, 2006.)

Neke od najpopularnijih metoda rudarenja podataka, od kojih ćemo neke detaljnije objasniti u nastavku ovog poglavlja su:

1. Određivanje najbližeg susjeda
2. Grupiranje
3. Asocijativna pravila
4. Stabla odlučivanja
5. Klasteriranje
6. Umjetne neuronske mreže
7. Genetički algoritmi

Postoji još puno njih koje nisu veoma popularne, ali svejedno neke od njih trebalo bi istaknuti, to su:

- Metode potrošačke košarice
- Memorijski temeljeno razlučivanje
- Bayesove mreže
- Neizrazita logika

Kao što je gore napisano da su najpopularnije metode, također to isto možemo vidjeti i u tablici 1. koja prikazuje učestalosti korištenja metoda rudarenja podataka. Tablica je napravljena prema podacima sa internetske stranice KDnuggets koja se bavi otkrivanjem znanja iz baza podataka. Iz tablice je vidljivo da su stablo odlučivanja, regresija i klasteriranje najpopularnije metode.

Metode rudarenja podataka učestalo korištene u posljednjih 12 mjeseci (N=203)

Stabla odlučivanja / pravila (127)	62.6%
Regresija (104)	51.2%
Klasteriranje (102)	50.2%
Deskriptivna statistika (94)	46.3%
Vizualizacija (66)	32.5%
Asocijativna pravila (53)	26.1%
Analiza vremenskih serija (35)	17.2%
Neuronske mreže (35)	17.2%

Tablica 1: Metode rudarenja podataka prema učestalosti korištenja

Sve metode imaju isti glavni cilj, a to je da prikažu kretanje podataka na temelju kojih možemo donijeti kvalitetne zaključke. Svaka današnja metoda je nastala kao rezultat dugotrajnog rada, razvoja te istraživanja statičkih algoritama koji se primjenjuju na sirovim podacima za otkrivanje znanja.

Prilikom provođenja rudarenja podataka preporuča se koristiti više od jedne metode ukoliko je to moguće. Rezultate dobivene različitim metodama možemo uspoređivati, a doneseni zaključci mogu biti bazirani na metodi koja nam najbolje odgovara.

3.4.1. Stablo odlučivanja

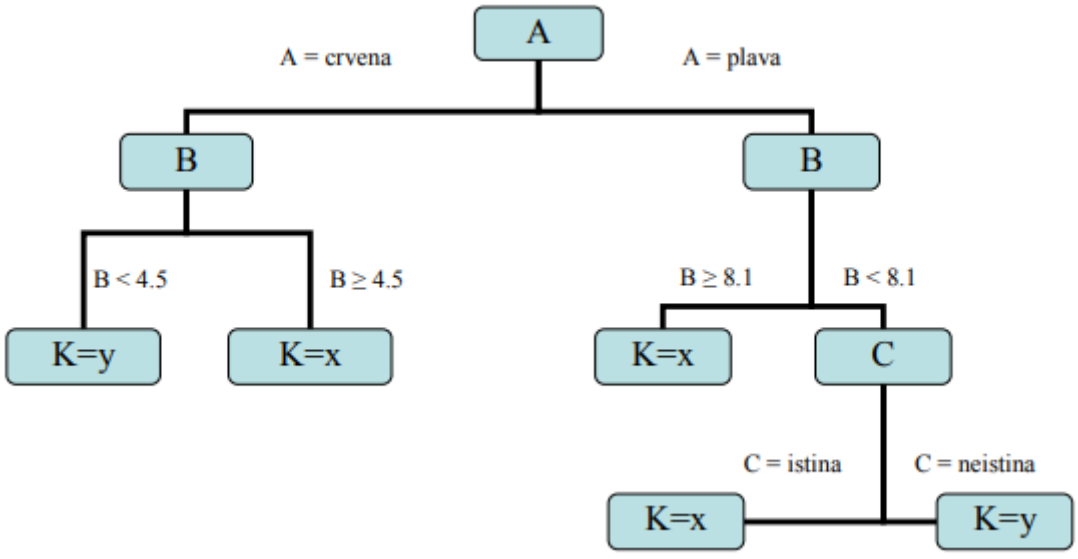
Ova metoda je vrlo dobra za klasifikaciju i predviđanje, ali i za procjene vrijednosti, klasteriranje, opisivanje i vizualizaciju. Naspram drugih metoda stablo odučavanja je veoma jednostavna i razumljiva metoda te je zbog toga veoma privlačna i popularna. Toliko je jednostavna da su njezina pravila napisana na čitljivom jeziku kojega svatko može pročitati te se direktno mogu koristiti u radu s bazama podataka i na taj način možemo određene primjere iz baze izdvojiti korištenjem pravila generiranih stablom odlučivanja. Također, ova metoda se

koristi i za istraživanja i uočavanja između veza velikog broja ulaznih varijabli prema traženoj vrijednosti. (Gamberger i Šmuc, 2001.)

Nastalo je 1730. godine kada je švedski botaničar Carl Linnaeus napravio prvo stablo odlučivanja u kojem je poznate žive stvari dijelio na kraljevstva, plemena, klase, staleže, obitelji, vrste i drugo. Nakon toga stablo odučavanja se veoma nadograđivalo te današnje stablo odlučivanja kako poznajemo je nastalo na bazi statističkih metoda raspoznavanja uzoraka.

Stablo odlučivanja koristimo kod podijele velike količine podataka u manje skupine kroz niz jednostavnih pravila. Također, pretvara skup podataka koji je raznolik, točnije heterogen u manje homogene skupine podataka koji imaju slične ili iste značajke. Osim toga, ovu metodu možemo koristiti i za izradu modela kada se dovršava neki sustav iako se koriste druge metode. Najčešće ovu metodu koristimo kada nam je sposobnost interpretacije modela od ključne vrijednosti. Uzmimo za primjer marketing gdje nam je potrebno dobro opisati različite segmente populacije kupaca za marketinške stručnjake kako bi oni mogli osmisliti efektivnu kampanju radi povećavanja prodaje određenih proizvoda.

Za konstruiranje stabla odlučivanja postoji velik broj različitih algoritama koji su veoma kvalitetni, ali najpoznatiji i vjerojatno algoritam koji se najviše koristi je C4.5 tj. njegova poboljšanja komercijalna verzija C5.0. Prvi algoritam ujedno i prethodnik C4.5 je algoritam pod nazivom ID3 te je prikazan sa slici 2. (Gamberger i Šmuc, 2001.)



Slika 4: Primjer jednostavnog stabla odlučivanja.

Algoritam ID3 pregledava sve atribute u skupu podataka te pronalazi onaj koji nabolje odvaja primjere u ciljne klase, ukoliko neki atribut savršeno odvaja klase, algoritam se zaustavlja. Ovaj algoritam koristi *„greedy“* (pohlepni) pristup, odnosno, traži trenutno najbolji atribut i nikada se ne vraća unatrag kako bi provjerio ispravnost prethodnih izbora. Moramo znati da ID3 je sklon pogreškama, točnije on može generirati stabla koja će nam dati pogrešne klasifikacije na skupu primjera za učenje. U konačnici, ovaj algoritam može generirati stablo dovoljno kompleksno da točno klasificira sve primjere iz skupa podataka za učenje. Vjerojatno vam to zvuči kao razumna strategija, ali ona donosi puno dodatnih problema, bilo zbog šuma u podacima ili nedovoljno velikog uzorka podataka. U oba ta slučaja, ID3 bi generirao stablo koje je *„pretjerano dobro“* klasificira primjere iz skupa za učenje na štetu svih ostalih. (Gamberger i Šmuc, 2001.)

Iako ga zovemo stablom odlučivanja ono ipak izgleda više kao korijenje drveta. Naše stablo sadrži početno čvorište(eng. *root node)* iz kojeg sve počinje te se ono grana na više manjih grana. Čvorišta koja izlaze iz početnog čvorišta nazivamo djeca čvorišta(eng. *rhild nodes)* te ukoliko iz njih postoji daljnje grananje tada ta čvorišta istovremeno i postaju roditelji čvorišta(eng. p*arent nodes).* Zadnje čvorište u kojem završava naše stablo naziva se završni čvor(eng. *leaf node)*. Naše varijable ulaze u početno čvorište te na temelju testa odlučuje se u koju od grana će otići varijabla. Ovi testovi se izvode na temelju različitih algoritama koji imaju isti cilj, odabrati test koji najbolje razlikuje klase kroz cijelo stablo te taj proces ponavljamo skroz dok ne dođemo do završnih čvorova. Od samog početka tj. početnog čvora pa se do završnog čvora postoji jedinstveni put, taj put je izraz pravila koja su se koristila za klasifikaciju varijabli. Svi slojevi stabla ne moraju imati isti broj čvorišta te zbog toga stablo u konačnici može poprimiti raznorazne oblike.

Kada imam gotovo stablo moramo napraviti korekciju koja se radi na način da se uspoređuju dobiveni rezultati kod grananja na pojedinim čvorovima uz pomoć nekih od testova kao što su C2, C5, Gini testovi i slični. Oni funkcioniraju tako da dokažu da je grananje dobro napravljeno te ako su razlike u rezultatima značajne ona moramo napraviti regrupaciju grananja i tako dobivamo novi izgled stabla. Ove testove možemo napraviti prije ili poslije iz jednog čvorišta.

Stablo odlučivanja može rasti beskonačno, ali isto tako može imati i premali broj grana koje nam nisu od nikakve koristi. Postoje dvije tehnike u određivanju kompleksnosti i dubine stabla kako bi postigli optimalni broj grana unutar stabla. Prva tehnika je da odozgo prema dolje

pomoću određenih pravila prekida daljnji rast stabla u određenom smjeru, dok druga tehnika je odozdo prema gore koja radi na način da odstranjuje grane sa stabla najveće kompleksnosti dok ne postignemo željenu složenost. Ove dvije tehnike najbolje je koristi zajedno kako bi postigli bolji i optimalniji rezultat u kreiranju stabla.

Prednosti metode stabla odlučivanja su:

- Relativno mali zahtjevi za računalnim resursima
- Sposobnost da koristimo sve tipove atributa
- Laka čitljivost
- Sposobnost generiranja razumljivih modela

Nedostatci metode stabla odlučivanja su:

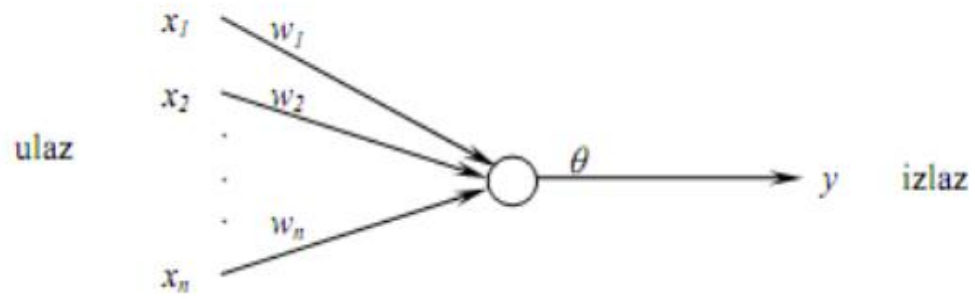
- Sklonost greškama u višeklasnim problemima sa relativno malim brojem primjera za učenje
- Nekada generiranje stabla može biti računalno zahtjevan problem
- Manja prikladnost za probleme kod kojih se traži predikcija kontinuiranih vrijednosti ciljanog atributa

(Gamberger i Šmuc, 2001.)

3.4.2. Neuronske mreže

Jedna je od najpopularnijih metoda današnjice zbog toga što se dokazala u mnogim aplikacijama za pomoć u odlučivanju. Ova metoda je moćna, usmjerena i brza za predviđanje, klasifikaciju i klasteriranje.

Neuronska mreža je sustav međusobno povezanih računskih elemenata te nalikuje povezanom usmjerenom grafu. One funkcioniraju na isti način kao i ljudski mozak ili neke druge biološke mreže. Osnovni element je neuron i on nalikuje čvoru grafa s pripadajućim ulaznim i izlaznim granama. Ideja neuronskih mreža je da skup dolaznih informacija iz više izvora u jednu jedinicu, točnije neuron kombinira u jednu izlaznu informaciju. Ulazni dio neurona čini realni vektor ulaznih vrijednosti (x_1, x_2, …, x_n) dok je izlazni dio sam jedna vrijednost y, pogledati sljedeću sliku.



Slika 5: Neuron, osnovni element neuronske mreže (Ujević, str. 81).

Velika važnost se pridodaje svakoj vrijednosti jer i ona jako mala utječe na promjenu konačnog ishoda. Kompleksnost i snaga ove metode baš proizlazi iz takvih uzoraka podataka. Nastanak neuronske mreže je takav da se neuroni međusobno povezuju na način da izlazna vrijednost jednoga bude ulazna vrijednost jednog ili više drugih neurona. Neuronska mreža može imati beskonačno ulaza, izlaza i skrivenih slojeva te sve ovisi o snazi računala na kojem se obrađuje. Ulaz u neuronsku mrežu čine neuroni sa slobodnim izlazima dok izlaz i nje čine neuroni sa slobodnim izlazima. (Ujević, str. 81, 2004.).

U nastavku ćemo pokušati bolje objasniti kako jedan neuron funkcionira kako bi mogli lakše shvatiti kako zapravo cijela mreža funkcionira. Sama struktura neuronske mreže je potpuno fleksibilna u odabiru funkcija koje obavlja. Neuron obavlja aktivacijsku funkciju, odnosno, kroz njega prolaze podaci i u ovisnosti o tim podacima određena funkcija se aktivira i obrađuje podatke na sebi svojstven način. Aktivacijsku funkciju dijelimo na dva dijela:

1. **Funkcija kombinacija** – ona spaja sve ulazne veličine u jednu izlaznu vrijednost. Preciznije, ova funkcija koristi jednostavne algoritme za obradu podataka, kao što su suma svih vrijednosti, uključivanje maksimalne ili minimalne vrijednosti ulaznih podataka te logičke operatore AND ili OR, sve te vrijednosti se mogu kombinirati, ali većinom je dovoljna samo jedna
2. **Funkcija prijenosa** – ona prenosi vrijednost od funkcije kombinacije na izlaz i neurona. Neke tipične funkcije prijenosa su: linearne funkcije, sigmoide, hiperbolične i druge.

Kako bi neuronske mreže bile što optimiziranije moramo ih trenirati. Treniranje neuronskih mreža zapravo je korištenje skupa za učenje te modificiranje težinskih vektora i pragove neurona kako bi pretraživali prostor rješenja optimiziranjem parametara klasifikacijskog modela. Kada trenirano našu mrežu, primjeri iz skupa za učenje se koriste jedan po jedan te se za svaki od njih računa izlazna vrijednost i uspoređuje za traženim izlazom. Ovisno o razlici između te dvije vrijednosti težinski vektor i vrijednost praga se korigiraju obrnuto proporcionalno veličini greške koju su uzrokovali u izlaznoj vrijednosti. Drugim riječima, ako je rezultat veći od traženog onda se smanjuju težine ulaza koje generiraju razliku i obrnuto. Prva i najčešća metoda treniranja naziva se propagacija greške unatrag(eng. *backpropagation*), ona koristi iterativan postupak za propagaciju greške tj. razlike stvarnog i traženog izlaznog, od neurona izlaznog sloja prema unutarnjim slojevima mreže. Korekcija odgovarajućih težinskih vrijednosti prati propagaciju greške te se na taj način korekcija parametara širi od izlaznog prema ulaznom sloju mreže te kada se sve težinske vrijednosti u neuronskoj mreži stabiliziraju ovaj algoritam staje. (Ujević, str. 82-83, 2004.).

U praksi se koristi puno različitih topologija neuronskih mreža. Također, kod klasifikacije neuronske mreže pokazale su se kao jako dobar izbor na težim klasifikacijskim problemima kod kojih je teško ili nemoguće koristiti klasične tehnike simboličkog učenja. Uz sve to, neuronske mreže su veoma prilagodljive u uvjetima šuma u podacima. (Ujević, str. 83, 2004.).

Međutim, postoje i neki nedostaci neuronskih mreža, a to su:

- Relativan spor i zahtjevan proces indukcije modela u usporedbi s klasičnim tehnikama
- Klasifikacijski modeli nisu eksplicitno izraženi u obliku strukturnog opisa važnih odnosa među varijablama
- Implicitni model nije razumljiv niti podložan verifikaciji ili interpretaciji

(Ujević, str. 83, 2004.).

3.4.3. Algoritam k srednjih vrijednosti

Za razliku od već objašnjenih metoda ova se razlikuje po tome što ne služi za klasifikaciju, nego za klasteriranje. Metode klasteriranja podataka spadaju u grupu neusmjerenih metoda kojima je cilj otkrivanje globalne strukture podataka. Ovaj pristup nema definirani ciljni atribut kao što ga ima u usmjerenim metodama te zbog toga ne postoji razlika između atributa. (Gamberger i Šmuc, 2001.)

Metode klasteriranja koristimo zbog podjele primjera u skup podskupova ili klastera koji moraju zadovoljavati dva osnovna kriterija, a to su:

- Svaki klaster mora predstavljati homogeni skup, odnosno, primjeri koji pripadaju istoj grupi su međusobno slični
- Svaki klaster se mora razlikovati od ostalih klastera, odnosno, primjeri koji pripadaju određenom klasteru se moraju značajno razlikovati od primjera koji pripadaju ostalim klasterima

(Gamberger i Šmuc, 2001.)

Klasteri mogu biti:

- Ekskluzivni – svaki primjer pripada isključivo jednom klasteru
- Preklapajući – jedan primjer može istovremene pripadati nekolicini klastera
- Probabilistički – svaki primjer pripada svakom od klastera s određenom vjerojatnosti
- Hijerarhijski strukturirani – grupa podjela primjera na najvišoj razini te finijom podjelom na nižoj razini

U slučajevima kada očekujemo postojanje klastera u podacima onda koristimo metode klasteriranja uz pomoć kojih otkrivamo klastere podataka koji bi trebali predstavljati skupove primjera koji imaju puno toga zajedničkog. Takav način stvaranja klastera nam može uvelike smanjiti kompleksnost određenog problema ukoliko ga koristimo prije primjena drugih metoda bilo to stabla odlučivanja, neuronske mreže ili nečeg drugoga. Također, dobivene podskupove možemo modelirati odvojeno te nam takav način na kraju može dati puno bolje rezultate u prediktivnom ili deskriptivnom smislu nego što bi ih dobili bez prethodne primjene metode klasteriranja podataka (Gamberger i Šmuc, 2001.)

Kao ulaznu vrijednost algoritam k srednjih vrijednosti koristi prethodno definirani broj klastera(k). U ovoj metodi nužno je uvesti pojam višedimenzionalnog prostora koji je definiran atributima kao osima tog prostora te srednja vrijednost se odnosi na prosječnu lokaciju u tom prostoru dok se vrijednost svakog atributa odnosi na udaljenosti tog primjera od ishodišta. Kako bi ove geometrija bila što efikasnija moramo koristiti numeričke vrijednosti atributa te sve vrijednosti nominalnih atributa trebaju se pretvoriti u numeričke i na kraju ih normalizirati da bi omogućili ravnopravno izračunavanje po svim koordinatama prostora (Gamberger i Šmuc, 2001.)

Ovaj algoritam je jednostavna i iterativna procedura u kojoj centralnu ulogu ima centorid, odnosno, umjetna točka u prostoru koja predstavlja srednju lokaciju određenog klastera primjera. Koordinate centroida izračunavaju se kao prosječne vrijednosti svih primjera koji pripadaju klasteru (Gamberger i Šmuc, 2001.).

Algoritam: algoritam k srednjih vrijednosti.
Ulaz: k - broj klastera.
Koraci:

1. Slučajno odabrati k točaka kao početne točke centroida svih k klastera. To mogu biti i primjeri iz skupa podataka.
2. Dodijeliti svaki primjer centroidu kojem je primjer najbliži, formirajući na taj način k ekskluzivnih klastera primjera.
3. Izračunati nove centroide klastera na način da se izračuna prosječna vrijednost, po pojedinim atributima, svih primjera koji pripadaju odreñenom klasteru.
4. Provjeriti jesu li centroidi klastera promijenili svoje koordinate više od prethodno definiranih minimalnih vrijednosti.
   Ako jesu, prijeći na točku 2.
   Ako nisu, određivanje klastera je završeno.

Slika 6: Algoritam k srednjih vrijednosti (Gamberger i Šmuc, 2001.)

Ukoliko odaberemo pogrešan broj klastera(k) u ovom algoritmu nećemo dobiti točne konačne rezultate. Način na koji ćemo dobiti točan broj klastera je da eksperimentiramo s različitim brojem klastera te odaberemo onaj za koji smatramo da je najtočniji. U biti, optimalan broj klastera će biti onaj broj koji najbolje odgovara omjeru intragrupnih i intergrupnih udaljenosti primjera. Bolje, jače i sofisticiranije metode klasteriranja taj broj automatski određuju (Gamberger i Šmuc, 2001.)

3.4.4. Asocijativna pravila

Ova metoda je veoma jasna i iskoristiva te ju zbog toga koristimo u analizi tzv. potrošačkih košarica. Asocijativna pravila jasno izražavaju u kojoj su mjeri važni proizvodi korelirani te tim sugeriraju konkretne akcije. Koriste se uglavnom u obradi podataka kod kojih su atributi nominalnog, odnosno, kategoričkog tipa.

Za učinkovitu primjernu ove tehnike moramo riješiti sljedeće probleme:

1. **Izbor pogodnog skupa elemenata** – detaljni podaci skupljeni na licu mjesta, većinom je to mjesto prodaje, su osnova za obradu podataka ovom metodom, ali to ne znači da ćemo te konkretne artikle u transakcijama uzeti za proces obrade podataka. Artikli u prodavaonicama su najčešće svrstani prema kategorijama, izbor prave razine kategorizacije može imati ključnu ulogu u smislenosti konačnih pravila, ali i u redukciji velikog broja artikala u jedan. Ponekad čak do stotine artikala se mogu svrstati u jedan koji će dobro reprezentirati generalna svojstva svih artikala u toj kategoriji.
2. **Veliki broj elemenata koji se pojavljuju u velikom broju interesantnih pravila** – iz razloga što broj kombinacija za skupove s više elemenata raste eksponencijalno s brojem elemenata u transakcijama, broj potrebnih izračuna mjera(značaj, pouzdanost, poboljšanje) skupova elemenata za velike trgovačke centre brzo poraste preko milijun. Objasnimo ovo na primjeru, recimo da imamo 1000 različitih artikala, a ukupan broj mogućih skupova od tri elementa je $\binom{1000}{3} = 166.167 \cdot 10^6$. Kao što iz ovoga možemo i zaključiti izračun frekvencija i mjera kvalitete za skupove s pet ili više elemenata je vrlo vjerojatno vremenski neizvedivo te je zbog toga vrlo važno koristiti taksonomiju, odnosno, generalizaciju elemenata.

Pošto se ova metoda najčešće koristi u obradi podataka u obliku transakcija moramo znati nekoliko termina u terminologiji asocijativnih pravila, a oni su:

- Element – u terminologiji rudarenja podataka to je par atribut-vrijednost
- Transakcija – u terminologiji rudarenja podataka to je primjer
- Skup transakcija - u terminologiji rudarenja podataka to je skup podataka

Najčešće transakcije se razlikuju u broju elemenata, što inače nije slučaj sa podacima koje koristimo kod drugih metoda modeliranja te zbog toga za većinu metoda modeliranja podataka nužno je pretvoriti transakcijske podatke.

U našem skupu transakcija svaka transakcija daje informaciju o tome koji se elementi zajedno pojavljuju u transakciji te je onda uz pomoć toga vrlo lako napraviti tablice koje prikazuju frekvenciju pojavljivanja parova. Nakon toga, iz napravljene tablice vrlo lako je napraviti jednostavna pravila, primjer jednostavnog pravila glasi ovako:

R_1 = „Element 1 se zajedno s elementom 2 pojavljuje u 20% svih transakcija“

Ovo pravilo nam govori da je 20% mjera frekvencije pojavljivanja para elemenata 1 i 2 u skupu svih transakcija te ono predstavlja signifikantnost pravila. Omjer broja transakcija u kojemu će se pojavljivati oba elementa sa brojem transakcija u kojem se pojavljuje samo element 1 naziva se pouzdanost. Ako uzmemo da je frekvencija pojavljivanja elementa 1 u svim transakcijama 15%, a frekvencija pojavljivanja elementa 2 iznosi 20% tada pouzdanost našeg pravila R_1 iznosi c(R_1) = 10/15 = 0.666.

Od napisanog pravila lako možemo napraviti i inverzno pravilo te će ono glasiti:

R_2 = „Element 2 se zajedno s elementom 1 pojavljuje u 20% svih transakcija“

Iako na prvu mislimo da se radi o isto pravilu ipak svojstava R_1 i R_2 se razlikuju. Iz razloga što je frekvencija pojavljivanja elementa 2 u cijelom skupu transakcija 20%, pouzdanost našeg pravila R_2 je c(R_1) = 10/20 = 0.5. Pouzdanost pravila R_2 zapravo nam govori da kada se u transakciji pojavi element 2, vjerojatnost da će se u istoj transakciji pojaviti i element 1 je 0.5 ili 50%.

Generiranje asocijativnih pravila je iterativan proces koji ide po sljedećoj shemi:

1. Generiranje tablice frekvencija pojavljivanja pojedinačnih elemenata.
2. Generiranje tablice frekvencija pojavljivanja dva različita elementa. Iz tablice se izdvajaju parovi s poboljšanjem većim od unaprijed zadanog kriterija.
3. Generiranje tablice frekvencija pojavljivanja tri različita elementa. Iz tablice se izdvajaju tripleti s poboljšanjem većim od unaprijed zadanog kriterija.
4. Itd…

(Gamberger i Šmuc, 2001)

Prednosti asocijativnih pravila su:

- Jednostavnost i jasnoća
- Metoda namijenjena problemima koji nisu klasifikacijskog tipa
- Može se koristiti kod primjera koji imaju varijabilni broj atributa
- Algoritmi za generiranje asocijativnih pravila su zapravo vrlo jednostavni

NAMJENA / METODA	Sumiranje podataka	Segmentacija	Klasifikacija	Predviđanje	Asocijacija (prepoznavanje uzoraka)
Deskriptivna statistika i vizualizacija	+	+			+
Korelacijska analiza					+
Klasteri		+			
Diskriminantna analiza			+		
Regresijska analiza				+	+
Neuronske mreže		+	+	+	+
CBR					+
Stabla odlučivanja			+	+	
Pravila asocijacije					+

Tablica 2:Koju metodu rudarenja podataka koristiti za pojedinu namjenu (Izvor: Zekić – Sušac, 2009.)

4. Primjena rudarenja podataka u upravljanju znanjem

Kao što već znamo pojavljivanje interneta i dostupnosti velike količine sadržaja iz našeg vlastitog doma dovelo je do masovnog povećanja podataka. Veliki dio tih podataka većina smatra beskorisnim, ali iz njih se može puno toga naučiti. Velike organizacije znaju temeljiti svoja poslovanja na tim podacima. Rudarenje podataka danas ima veliku primjenu u upravljanju znanja spomenimo da je Američka vojska uz pomoć nje uspjela identificirati vođu napada na Twin Towers za koju smo zasigurno svi čuli, a također i danas ju koriste CIA i Canadian Security Intelligence Service. U ovom poglavlju izdvojiti ćemo jedan stvarni primjer te ga detaljno objasniti, a radi se o primjeni rudarenja podataka u bankarstvu.

4.1. Primjena rudarenja podataka u bankarstvu

4.1.1. Rizik

Upravo naslov ovog djela je tipičan za banke ili osiguravajuća društva zbog toga što je bankama veoma važno da ne daju kredite osobama koje nisu u mogućnosti vratiti ih dok je osiguravajućim društvima rizik da će klijenti iskoristiti osiguranje zbog recimo nekakve ozljede. Banke u ovom slučaju koriste napravljene modele za predviđanje hoće li klijent moći vratiti krediti ili ne. Ovakve modele moguće je koristiti i za klasične kredite koji imaju neki oblik osiguranja, ali i za neosigurane kredite. Osim ovog rizika, rizik od prijevare je također važan za banke i osiguravajuća društva jer na primjer kod krađe kreditnih kartica banke preuzimaju dio štete na sebe. Iz tog razloga banke izrađuju modele na temelju ponašanja kupaca te uz pomoć njega mogu brzo otkriti radi li se o krađi ili ne te tako smanjuju gubitke banke. S druge strane osiguravajuća društva imaju rizik da će klijenti pokušati dobiti osiguranje na temelju prijevare, na primjer da sami podmetnu požar. Zbog toga oni izrađuju modele koji im olakšavaju detekciju pokušaja prijevare te modele koji će predviđati hoće li klijent u budućnosti pokušati prevariti osiguravajuće društvo.

4.1.2. Prodaja dodatnih proizvoda postojećim klijentima

Kao što i sam naslov kaže banke uz pomoć svojih modela prodaje dodatnih proizvodima postojećim kupcima određuje vjerojatnost da će klijenti kupiti i neke dodatne proizvode. Cilj ove analize nije samo nagovaranje klijenta da kupi dodatne stvari nego ponudom proizvoda odabranim klijentima povećava se kvaliteta odnosa s njim. Zbog toga raste profitabilnost poslovanja zato što trošak prodaje već postojećim klijentima puno je manji nego privlačenje novih klijenata te također još i povećavamo lojalnost postojećih klijenata.

4.1.3. Zadržavanje postojećih klijenata

Odlazak postojećih klijenata drugim bankama, odnosno, konkurenciji nije samo problem koji se veže uz banke nego je problem i mnogih drugih djelatnosti. Zbog toga što je današnje tržište zasićeno jedina mogućnost rasta organizacija je preotimanje klijenata od konkurencije ili prodaju drugih proizvoda postojećim klijentima. Zbog bolje ponude konkurencije klijenti često prelaze iz jedne u druge banke. Već dugi niz godina kartične kompanije vode rat kamatama kako bi pridobile što više klijenata te tako u početku nude niske kamate i nadaju se da će im klijenti ostati i nakon isteka tih pogodnosti, ali često to nije slučaj. Istraživanja u svijetu pokazala su da klijenti znaju vješto koristiti pogodnosti koje im se pružaju te nakon njihovog isteka prijeći konkurenciji koja nudi bolju ponudu. Iz tog razloga uz pomoć rudarenja podataka se izrađuju modeli koji mogu predvidjeti hoće li klijent, nakon što se kamate podignu, prijeći konkurenciji ili smanjiti potrošnju.

4.1.4. Segmentacija

Glavni resurs banke su njezini klijenti te na temelju poznavanja njihovih karakteristika, preferencija i specifičnih potreba banka im može prilagoditi ponudu svojih usluga, ali pri tome treba uzeti u obzir da se karakteristike klijenata mijenjaju s godinama. Objasnimo to na primjeru jedne osobe, većina ljudi otvara bankarski račun na početku ili za vrijeme studiranja te kao novi korisnik koristi jedan ili tek nekoliko usluga. Ista ta osoba nakon nekoliko godina kada se zaposli koristiti će neke druge usluge, a također kada se umirovi će koristiti opet drugačije usluge. Banke skupljaju velike količine o klijentima te ih koriste za analizu karakteristika klijenata i na temelju njih formiraju segmente kojima se mogu posebno prilagoditi usluge. Već godinama banke koriste tradicionalne segmentacije sektora stanovništva i poduzeća, ali one često znaju prikazati pogrešno stanje. Uz pomoć rudarenja podataka banke mogu pronaći segmente koje su prije zanemarivale te uz pomoć njih mogu ponuditi prilagođene proizvode koji će donijeti povećanju profitabilnosti poslovanja.

4.1.5. Životna vrijednost klijenata

Što nam zapravo ovaj naslov govori? Životna vrijednost klijenata je vrijednost koju očekujemo od pojedinog klijenta kroz određeno vrijeme. Zbog toga možemo zaključiti da je bankama u cilju privući što više studenata od kojih će veliki broj postati profitabilni klijenti. U početku zarada od studenata je mala, ali s vremenom ako banka stvori dobar odnos s njima može u budućnosti ostvariti veliku korist. Nakon što student diplomira velika vjerojatnost je da će trebati kredit za stan ili auto te će mu sigurno trebati tekući račun, kreditne kartice, životno

osiguranje itd. Zbog toga što pohađa fakultet ta osoba će biti visoko obrazovana te se od nje očekuje da će imati primanja iznad prosjeka te će si moći priuštiti više proizvoda. Rudarenjem nad podacima izrađuju se modeli koji predviđaju životnu vrijednost klijenata kako bi se bankarski djelatnici mogli više posvetiti koji trenutno nisu profitabilni, ali bi mogli biti u budućnosti.

4.1.6. Odaziv

Cilj ovog modela je predvidjeti koji će kupci pozitivno odgovoriti na ponudu kupovine proizvoda ili usluga, pri čemu se najčešće radi o direktnom marketingu. Poruku klijentima možemo poslati na različite načine, neki od njih su pošta, telefon ili putem interneta. S ovim modelom želimo privući nove kupce, ali i stare s kojima duže vrijeme ne poslujemo jer takve kupce ćemo lakše pridobiti na kupovinu proizvoda.

4.1.7. Aktivacija

Ovaj model nam predviđa vjerojatnost hoće li klijent kojeg smo pridobili postati profitabilan. Uzmimo za primjer da klijent uzme kod nas ugovor za životno osiguranje nakon čega ne uplaćuje premiju ili da uzme kreditnu karticu, ali ju ne koristi. Uz pomoć ovog modela želimo izbjeći situacije slične navedenom tako da klijentima ponudimo dodatne pogodnosti da bi ih potakli na aktivaciju ili jednostavno možemo odustati poslovati s njima.

4.1.8. Racionalizacija poslovanja

Uz pomoć rudarenja podataka možemo racionalizirati poslovanje kako bi ostvarili znatne uštede. Objasniti ću ovaj dio kroz nekoliko primjera. Kao što znamo sve banke danas posjeduju bankomate iako je to veliki izazov za banke u organizacijskom i logističkom smislu. Tehnički bankomati mogu držati veliku količinu novca, ali iz ekonomske perspektive nema smisla puniti sve bankomate do vrha jer je dnevni promet na bankomatima puno manji. Osim toga, novac na bankomatima ne donosi nikakvu profit. Pomoću rudarenja podataka možemo izraditi sustav za optimizaciju upravljanja gotovine koji bi mogao predviditi kada i koliko novca je potrebno isporučiti te bi pri tome uzimao u obzir tjedne, mjesečne i godišnje oscilacije. Također, može se koristiti i za izradu modela koji davao preporuku što učiniti kada klijent kasni s obročnim plaćanjem kredita ili premije osiguranja. Čak nekoliko banaka je koristilo rudarenje podataka u sigurnosne svrhe, odnosno, modelom su analizirali poslovnice koje su nedavno bile opljačkane kako bi tamo poslali jače osiguranje.

(Pejić-Bach: Rudarenje podataka u bankarstvu, 2005.)

5. Primjer rudarenja podataka kao metoda upravljanja znanjem na stvarnim podacima

U današnje vrijeme na internetu postoje različiti podaci koji su javno dostupni, odnosno, svi ih mogu koristiti bez ikakvih plaćanja, kazni ili ostalih sankcija. Također, postoje mnogi različiti repozitoriji koji imaju niz različitih podataka svrstanih u baze podataka ili datoteke različitih formata. Jedan od takvih je i stranica kaggle.com s koje sam su preuzeti podaci za ovaj projekt. Nakon dugo vremena razmišljanja i istraživanja različitih skupova podataka odlučeno je uzeti skup podataka vezan uz nasilje oružjem na području Sjedinjenih Američkih državama o kojem ćemo pisati u nastavku ovog rada.

5.1. Opis problema

U ovom radu pokušati ćemo opisati kako i koje vrijednosti mogu utjecati na nasilje oružjem te koje posljedice možemo očekivati. Također, pokušati ćemo donijeti zaključke na temelju analize podataka koji će se odnositi na to kako spriječiti nasilje oružjem te ukoliko dođe do njega kako smanjiti težine posljedica. Zbog razloga što danas oružje postaje sve dostupnije pogotovo u većim zemljama poput SAD-a nužno je poduzeti mjere sigurnosti da bi se zaštitili od svakog oblika posljedica upotrebe njime. S ovim rješenjem i krajnjim zaključcima moglo bi se pomoći državnim upravama donijeti određene mjere i zakone kako bi se nasilje oružjem smanjilo, a idealni cilj bi bio da više ne postoji taj problem. Konačni cilj ovog projekta je smanjiti nasilje oružjem te što bi državne službe, ali i obično stanovništvo trebalo učiniti da spriječi nastanak nasilja oružja, ali ukoliko dođe do njega što trebaju učiniti kako bi maksimalno smanjili posljedice.

5.2. Alati

5.2.1. BigML

BigML je potrošna, programibilna i skalabilna platforma za strojno učenje koja olakšava rješavanje i automatizaciju klasifikacije, regresije, predviđanja vremenske serije, analize klastera, detekcije anomalije, otkrivanja udruženja i zadataka modeliranja uz pojedine teme. BigML pomaže tisućama analitičara, programerima softvera i znanstvenicima širom svijeta da riješe zadatke strojnog učenja *"end to end"*, neprimjetno pretvarajući podatke u djelotvorne modele koji se koriste kao udaljene usluge ili, lokalno, ugrađuju u aplikacije da bi se napravila predikcija.

Napravljen je u siječnju 2011. godine s ciljem da strojno učenje bude lako i lijepo za svakoga. Nakon nekoliko godina mukotrpnog rada stručnjaci iz BigML-a napravili su rješenje koje danas koriste različite organizacije svih veličina. Također, uspjeli su izgraditi sofisticirana rješenja temeljena na strojnom učenju koja izdaju prediktivne obrasce iz svojih podataka.

5.2.2. Kaggle

Kaggle je platforma za natjecanja prediktivnog modeliranja i analitike u kojima se statističari i rudari podataka natječu za izradu najboljih modela za predviđanje i opisivanje skupova podataka koje nabavljaju od tvrtki ili korisnika. U ožujku 2017. godine Google objavljuje da kupuje Kaggle te da će se oni pridružiti timu Google Cloud i nastaviti biti različiti brand.

5.3. Opis skupa podataka

Kao što je već napisano ovaj skup podataka skinut je besplatno sa stranice kaggle.com koja je navedena točno u literaturi. Ovaj skup podataka je javan te je njegovo korištenje dostupno svima. Kontekst ovog skupa podataka kako navodi njegov autor je manjak velike i lako dostupne količine detaljnih podataka o nasilju oružjem.

5.3.1. Sadržaj skupa podataka

Ovaj skup podataka bilježi preko 260 tisuća incidenata nasilja sa pištoljima s detaljnim informacijama o svakom incidentu te je to sve dostupno u datoteci formata CSV. Autor ovog skupa navodi da želi olakšati znanstvenicima i statističarima podataka proučavanje nasilje oružja i da daju informirana predviđanja o budućim trendovima. Ova datoteka sadrži podatke o svim zabilježenim incidentima nasilja oružja na području SAD-a između siječnja 2013 i ožujka 2018 godine.

5.3.2. Popis skupa podataka

Kao što je već rečeno, ovaj skup podataka se sastoji od preko 260 tisuća podataka raspoređen u 29 varijabli. Sve te varijable ćemo ukratko opisati u sljedećoj tablici, a neke od njih ćemo još detaljno objasniti tokom rada.

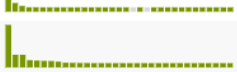
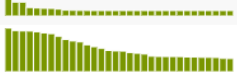
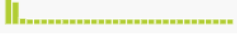
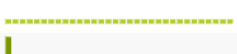
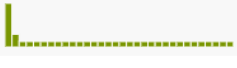
Originalni naziv	**Hrvatski naziv**	**Opis**	**Tip**	**Broj zapisa**
Incident_id	Incidenti_id	Id kaznenog izvješća	Numerički	239,677
Date	Datum	Datum zločina	Vremenski	239,677
State	Država	Država zločina	Kategorijski	239,677
City_or_county	Grad_ili_županija	Grad/županija zločina	Tekstualni	239,677
Address	Adresa	Adresa lokacija zločina	Tekstualni	223,180
N_killed	N_ubijenih	Broj ubijenih ljudi	Numerički	239,677
N_injured	N_ozlijeđenih	Broj ozlijeđenih ljudi	Numerički	239,677
Incident_url	Incident_url	URL u vezi s incidentom	Tekstualni	239,677
Source_url	Izvor_url	Referenca na izvor izvješća	Tekstualni	239,677
Incident_url_fields_missing	Incident_url_polja_nedostaju	True(istina) ako je incident_url prisutan, inače false(laž)	Kategorijski	239,677
Congressional_district	Kongresni_okrug	Okrug kongresne četvrti ID	Numerički	227,733
Gun_stolen	Pištolj_ukraden	Status oružja koji su uključeni u zločin(nepoznat, ukraden, itd.)	Kategorijski	140,179
Gun_type	Pištolj_vrsta	Tipizacija oružja	Tekstualni	140,226

		korištenih u zločinu		
Incident_characteristics	Incident_karatkeristike	Karakteristike incidencije	Tekstualni	239,351
Latitude	Širina	Mjesto događaja	Numerički	231,754
Location_description	Lokacija_opis	Nema opisa	Tekstualni	42,089
Longitude	Dužina	Mjesto događaja	Numerički	231,754
N_guns_involved	N_pištolja_sudjelovalo	Broj oružja uključenih u incident	Numerički	140,226
Notes	Bilješke	Dodatne informacije o zločinu	Tekstualni	158,647
Participant_age	Učesnik_godine	Starost sudionika u vrijeme zločina	Tekstualni	147,379
Participant_age_group	Učesnik _godine_grupa	Dobna skupina sudionika u vrijeme zločina	Kategorijski	197,558
Participant_gender	Učesnik_spol	Spol sudionika	Kategorijski	203,315
Participant_name	Učesnik_ime	Ime sudionika uključenog u zločin	Tekstualni	117,424
Participant_relationship	Učesnik_odnos	Odnos sudionika na druge sudionike	Kategorijski	15,774
Participant_status	Učesnik_status	Širina šteke učinjena sudioniku	Tekstualni	212,051
Participant_type	Učesnik_tip	Vrsta sudionika	Kategorijski	214,814
Sources	Izvori	Izvori podataka	Tekstualni	239,068

State_house_district	Država_kuća_okrug	Nema opisa	Tekstualni	200,905
State_senate_district	Država_senat_okrug	Nema opisa	Tekstualni	207,342

Tablica 3: Popis skupa podataka

Kada od cijelog ovog izvora podataka napravimo skup podataka u alatu BigML u zadnjem stupcu možemo vidjeti histogram svih redaka te uz pomoć njega možemo pronaći koje podatke je potrebno izbaciti da bi nam rezultat bio što točniji. Također, ovaj alat nam i sam prepoznaje neke retke koje je potrebno izbaciti te iz označuje sa crvenim uskličnikom. Sve je ovo vidljivo na slici 6. Osim napisanoga, ovaj alat čak i sam izrađuje određene retke odnosno, atribute kako bi nam olakšao i poboljšao naše konačne rezultate. U ovom slučaj on je od atributa *„Date“* napravio četiri dodatna atributa koji se odnose na godinu, mjesec, tjedan i dan, točnije napravio je atribute *„date.year“, „date.month-of-month“, „date.day-of-week“* i *„date.day“*. Također, on nam je automatski još označio atributa *„Date“* da je sada nepotreban jer imamo ove preostale točnije atribute.

Name	Type	Count	Missing	Errors	Histogram
incident_id	123	239,677	0	0	
date	YYYY-MM-DD	239,677	0	0	
state	ABC	239,677	0	0	
city_or_county	text	239,677	0	0	
address	text	223,180	16,497	0	
n_killed	123	239,677	0	0	
n_injured	123	239,677	0	0	
incident_url	text	239,677	0	0	
source_url	text	239,209	468	0	
incident_url_fields_missing	ABC	239,677	0	0	
congressional_district	123	227,733	11,944	0	
gun_stolen	ABC	140,179	99,498	0	
gun_type	text	140,226	99,451	0	
incident_characteristics	text	239,351	326	0	
latitude	123	231,754	7,923	0	
location_description	text	42,089	197,588	0	
longitude	123	231,754	7,923	0	
n_guns_involved	123	140,226	99,451	0	
notes	text	158,647	81,030	0	
participant_age	text	147,379	92,298	0	
participant_age_group	ABC	197,558	42,119	0	
participant_gender	ABC	203,315	36,362	0	
participant_name	items	117,424	122,253	0	
participant_relationship	ABC	15,774	223,903	0	
participant_status	text	212,051	27,626	0	
participant_type	ABC	214,814	24,863	0	
sources	text	239,068	609	0	
state_house_district	123	200,905	38,772	0	
state_senate_district	123	207,342	32,335	0	
date.year	YYYY-MM-DD	239,677	0	0	
date.month	YYYY-MM-DD	239,677	0	0	
date.day-of-month	YYYY-MM-DD	239,677	0	0	
date.day-of-week	M T W T F S S	239,677	0	0	

Slika 7: Početni skup podataka

Spomenut je već histogram. Histogram je nam grafički prikaz vrijednosti pojedinog atributa. Ukoliko nam je grafički prikaz ujednačen, odnosno, ne događaju se nikakve razlike velika je vjerojatnost da nam taj atribut nije potreban i trebamo ga izbaciti iz našeg skupa podataka. Razlog tome je to da ukoliko je vrijednost atributa uvijek ista za svaku instancu onda znači da ona ne donosi nikakvu promjenu u našim rezultatima. U ovom slučaju uzmimo atribut *„incident_url_fields_missing“* možemo vidjeti da je njegov histogram potpuno jednak cijelim svojim dijelom. Drugim riječima, vrijednost ovog atributa je *„false“* za svaku instancu našeg skupa te nam on neće činiti nikakvu razliku u konačnom rezultatu te ću ga zasigurno izbaciti iz skupa podataka.

Postoji još nekoliko atributa koje je potrebno izbaciti iz ovog skupa kako bi dobili u konačnici što bolje rezultate te iz njih donijeli što bolje zaključke. Atribut *„incident_id“* nam definitivno nije potreban jer on samo označuje svaku instancu različitim brojem i nema nikakav utjecaj. *„Incident_url“* nam također nije potreban jer to je atribut koji sadrži linkove, odnosno, poveznice na većinom novinske članke koje su pisale o tom događaju. Još jedan atribut koji je potrebno izbaciti je *„participant_age_group“* zbog toga što dobnu skupinu možemo iščitati i iz podataka od godinama sudionika te bi na taj način imali dvostruke podatke što nikako nije dobro za rudarenje. Atribut *„participant_type“* ćemo isto izbaciti zato što nam njegovi rezultati pokazuju samo da je uvijek *„Subject-Suspect“* što smo mogli i pretpostaviti jer više nitko ni ne može sudjelovati u zločinu.

Vjerojatno se pitate zašto smo neke atribute preskočili, a također ih je potrebno izbaciti. Razlog tome je taj što smo njih htjeli malo detaljnije objasniti te ih staviti u novi odlomak kako bi naglasili neke bitne stvari oko njih. Počnimo sa atributom *„gun_stolen“,* ovaj atribut može biti od velike koristi ukoliko posjedujemo dobre podatke za njega. U našem slučaju, autor koji je skupljao podatke iz nekog razloga nije mogao dobiti važeće podatke te je on postavio vrijednost svake instance tog atributa na *„Not stolen“.* Taj atribut mogao bi pokazivati kolike su vjerojatnosti da će osoba počiniti zločin ukoliko ukrade oružje te s druge strane kolike su šanse da netko bude ozlijeđen ili ubijen s ukradenim pištoljem. Ovo su samo neke moje pretpostavke za što bi mogao služiti, a sigurno postoji još puno razloga zašto bi ga koristiti te pretpostavljam da je ovaj atribut autor ostavio baš zbog toga razloga, odnosno, ukoliko uspije skupiti podatke i za taj atribut. U ovom slučaju ja sam ga izbacio iz skupa podataka jer kao što sam već i rekao, vrijednosti za ovaj atribut su u svakoj instanci jednaki te bi nam taj atributa loše utjecao na konačni ishod. Sljedeći atribut koji smo htjeli dodatno spomenuti je *„participant_releationship“*. Također, atribut koji bi mogao biti od velikog značaja u

budućnosti ovog skupa podataka, ali trenutno ne posjedujemo dovoljno podataka za taj atribut da bi ga koristili. U alatu BigML možemo vidjeti da imamo stupce *„Count“* i *„Missing“* koji predstavljaju koliko instanci posjeduje pojedini atribut, odnosno, koliko ih fali. Za ovaj atribut možemo vidjeti da samo 15,774 instanci od ukupno 239,677 što je veoma malo te ga iz tog razloga izbacujemo iz našeg skupa podataka. Zadnji atribut koji smo ostavili je *„State“* iz razloga što ga je BigML označio kao atribut koji je potrebno izbaciti, ali mi to nećemo učiniti. Nećemo ga izbaciti iz tog razloga što mislim da uz pomoć njega možemo donijeti neke velike i kvalitetne zaključke jer nam on govori u kojoj državi se dogodio zločin. S druge strane, ovaj alat nije pokazao da treba izbaciti atribute *„state_house_distirct“ i „state_senate_district“,* ali mi ćemo ih izbaciti jer ne shvaćamo dovoljno njihova značenja, a u opisu skupa podataka ne postoji opis za ta dva atributa.

Kao što je gore napisano i u poglavlju o rudarenju podataka, najbitniji i najdugotrajniji korak je prikupljanje i obrada podataka. Također, napisano je da je rudarenje podataka iterativan postupak te ukoliko smatramo da nešto nije u redu te da nam rezultati nisu dobri možemo se uvijek vratiti nekoliko koraka unazad i popraviti što je potrebno. Smatramo da je ovaj skup podataka trenutno kvalitetan te da ćemo iz njega moći dobiti kvalitetne i dobre rezultate kako bi mogli donijeti takve i zaključke. Ukoliko nešto ne bude odgovaralo u našim rezultatima, mi ćemo se također vratiti nekoliko koraka nazad te pokušati otkriti i popraviti dio u kojem sam pogriješili.

5.4. Klaster analiza nad stvarnim podacima

Kada govorimo o klaster analizi prvenstveno mislimo na grupiranje objekata na temelju sličnih karakteristika koje isti posjeduju. Razlika između klaster analize i faktorske analize je ta što klaster analiza koristi objekte kao predmet analize dok faktorska analiza koristi varijable, ali treba znati da se osobine objekata definiraju pomoću varijabli. Kod klaster analize koristimo podatke zadane od strane istraživača. Kod odabira varijabli vrlo je važno odrediti varijable koje najbolje reprezentiraju sličnost koja se istražuje prema tome ne možemo odabrati bilo koju varijablu. Na početku ne znamo niti broj klastera niti broj instanci koje istome pripadaju. Na kraju istraživanja instance su raspodijeljene u klastere u skladu s definiranim ciljem.

Metoda koju smo mi koristili kod analize i prikaza našeg skupa podataka je k-srednjih vrijednosti metoda. To je jedna od najpopularnijih metoda za analizu klastera u rudarenju podataka. Ona je jedna od metoda kvantiziranja vektora koja ima za cilj podijeliti *n* instanci u *k* klastera u kojima svaka instanca pripada skupini koja je njoj najsličnija po određenim brojčanim vrijednostima. Jedna od negativnih strana odabranog algoritma je ta što se za njega moraju koristiti isključivo numerički podaci.

5.4.1. Moja analiza

Nad već objašnjenim podacima prvo smo htjeli napraviti klaster analizu tako da nam sljedeće metode budu olakšane jer uz pomoć nje kasnije možemo dobiti točnije rezultate. Kao što je već objašnjeno u poglavlju na temu *„Rudarenje podataka“* metoda k-srednjih vrijednosti se radi tako da mi odaberemo neki k tj. broj klastera te nam alat prema tome odredi slične podatke i svrstava ih u klastere. U našem primjeru pokušali sam sa puno različitih brojeva klastera, uzimali smo prvo one najmanje te postepeno povećavali i onda odlučivali za koji mislimo da je najtočniji. Vrednovanje klaster analize je veoma teško te ne postoje metode ili načini kako odlučiti je li klaster analiza uspješna ili ne. Za naš primjer odlučili sam se za napraviti 8 klastera jer je to po našem mišljenju najtočnija vrijednost. Kako izgledaju podaci svrstani prema tim klasterima možete vidjeti na sljedećoj slici.

Slika 8: Klaster analiza

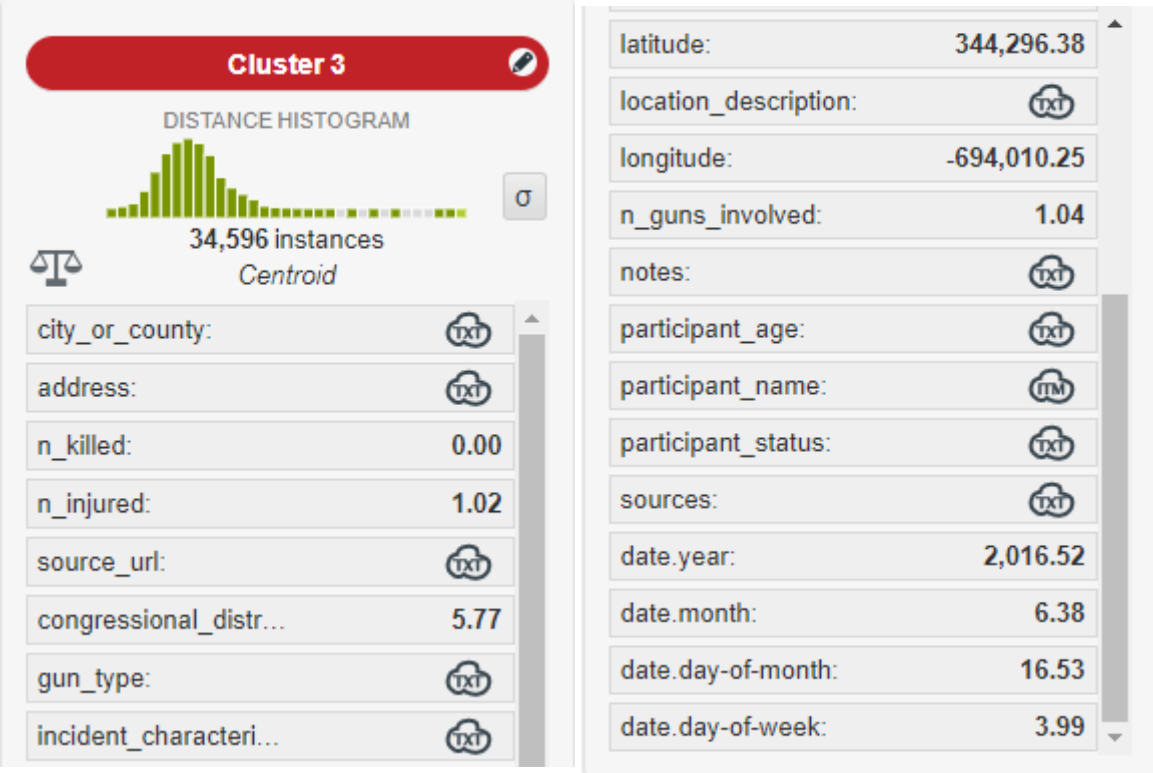
Osim izgleda klastera u ovom alatu možemo i dobiti izvješće sažetka klaster analize. Pogledajmo sljedeću sliku.

```
K-means Cluster (k=8) with 8 centroids
Data distribution:
    Global: 100% (128291 instances)
    Cluster 0: 12.66% (16239 instances)
    Cluster 1: 15.85% (20330 instances)
    Cluster 2: 17.76% (22782 instances)
    Cluster 3: 26.97% (34596 instances)
    Cluster 4: 17.77% (22798 instances)
    Cluster 5: 0.02% (31 instances)
    Cluster 6: 5.33% (6835 instances)
    Cluster 7: 3.65% (4680 instances)
```

Slika 9: Izvješće sažetka klaster analize

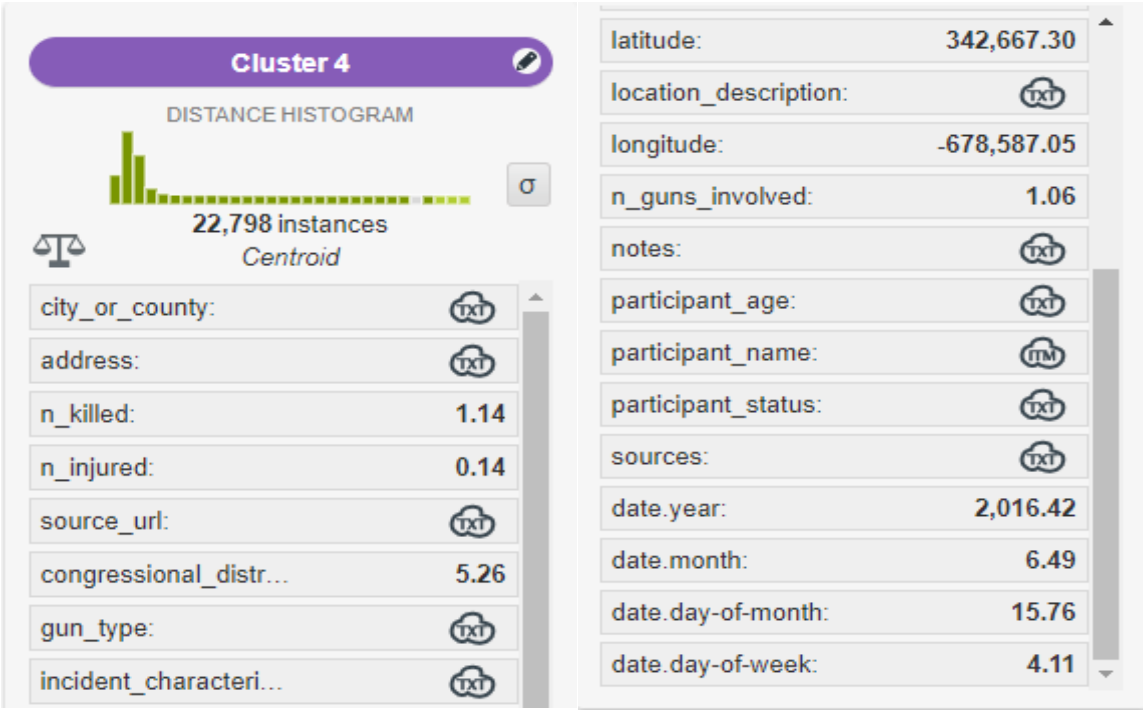
Ova slika prikazuje kako su se podaci distribuirali među klasterima. Točnije, možemo vidjeti da je klaster 5 najmanji klaster koji ima samo 31 instancu te iz toga možemo zaključiti da su to ili ekstremne vrijednost ili unos pogrešnih podataka te u oba slučaja te je podatke trebalo izbaciti iz skupa. Također, taj klaster se nalazi na sredini na slici 7 te je označen smeđom bojom i možemo vidjeti da svi ostale klaster su veoma blizu njemu što znači da ipak ima velikih sličnosti sa ostalim klasterima i zbog toga razloga bi te podatke trebalo sadržati, ali ta je odluka

na nama. Najveći klaster je na našoj slici označen crvenom bojom te nosi naziv klaster 3 i sadrži 34,596 instanci što čini čak 26.97% našeg skupa. U nastavku ovog poglavlja nastaviti ćemo objašnjavati svaki klaster zasebno, a pošto smo već djelomično započeli objašnjavanje klastera 3 onda ćemo od njega i krenuti te ići prema manjim klasterima.



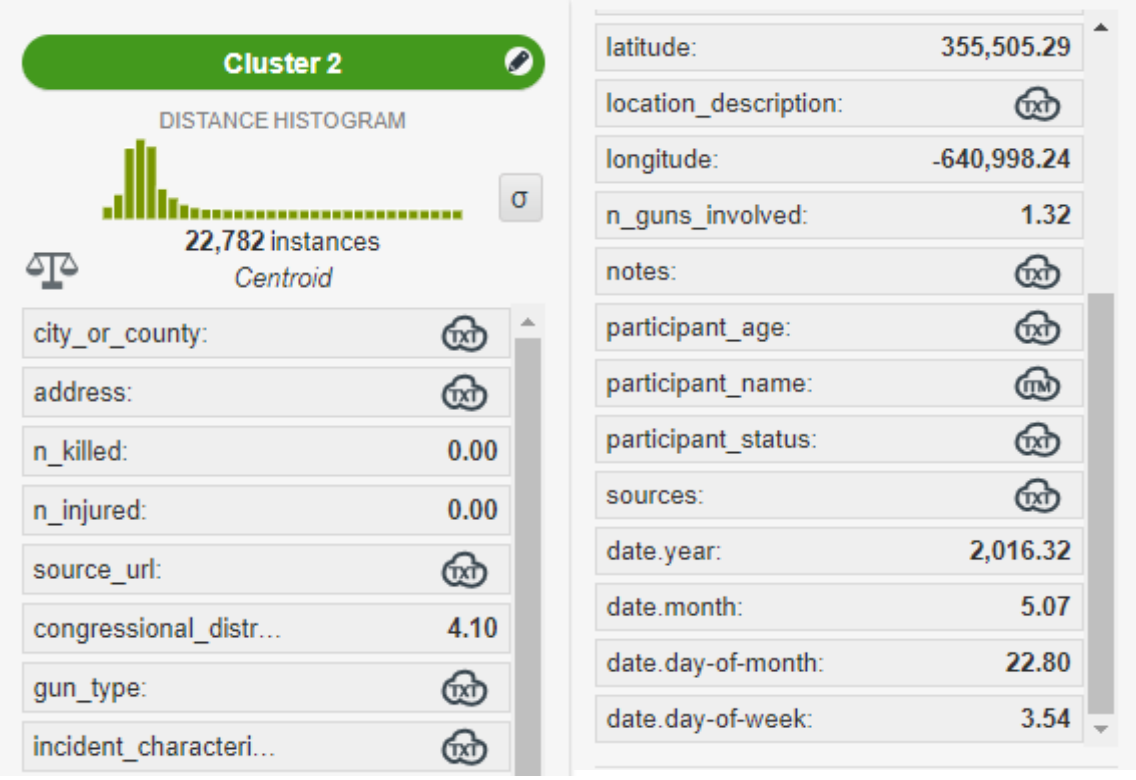
Slika 10: Podaci klastera 3

Iz gore dostupnih podataka možemo vidjeti da svim instancama je zajedničko da je broj ubijenih jednak nuli, odnosno, nema smrtno stradalih osoba. Također, broj ozljeđenih je malo veći od jedinice što znači da nije bilo ni puno ožljeđenima pri ovim zločinima i možemo zaključiti da ovaj klaster pokazuje događaje u kojima nije bilo velikog nasilja oružjem. Uz sve to vidimo da je i atribut koji se odnosi na broj uključenih pištolja malo veći od jedan pa iz toga bi mogli zaključiti da je u svim ovim zločinima samo jedna osoba bila napadač. Sljedeći klaster je pod brojem 4.



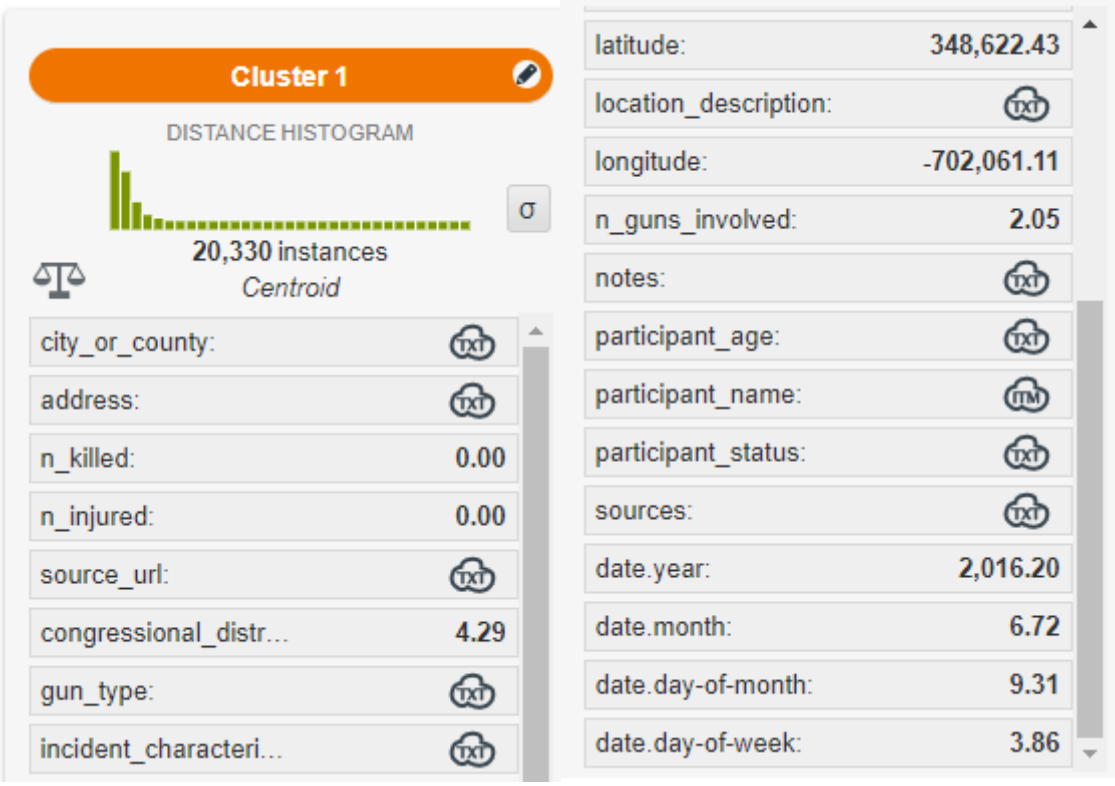
Slika 11: Podaci klastera 4

Iz podataka o klasteru četiri prvo što vidimo da broj ubijenih je bio veći od jedan te broj ozlijeđenih je malo veći od nule iz čega možemo zaključiti da se u ovim instancama oružje definitivno koristilo te je najmanje jedna osoba smrtno stradala. Također, vidimo da je broj uključenih pištolja sličan kao i kod klastera 3 i možemo reći da je tu također sudjelovao samo jedan napadač. Nakon ovog klastera dolazi nam klaster 2.



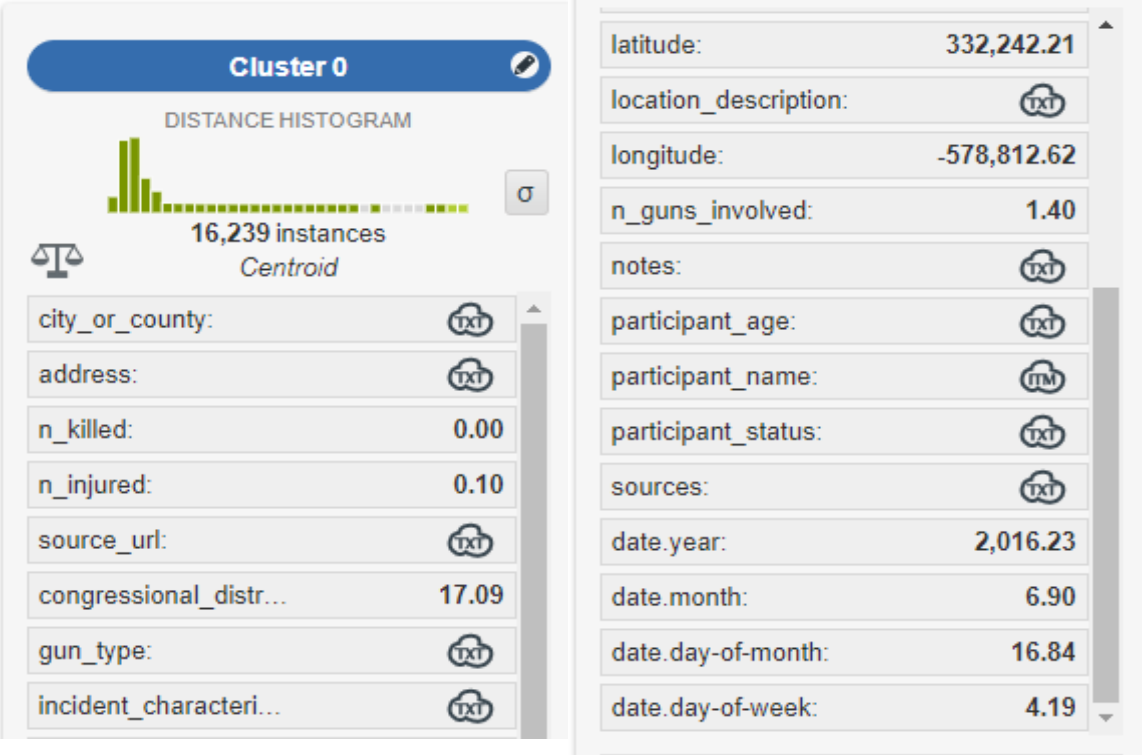
Slika 12: Podaci klastera 2

Ovdje možemo vidjeti da broj ubijenih i broj ozljeđenih je točno nula te da je broj korištenih pištolja dosta veći od jedan i znajući to možemo zaključiti da u ovim zločinima pištolj ili više njih je bio prisutan, ali nije bio iskorišten.



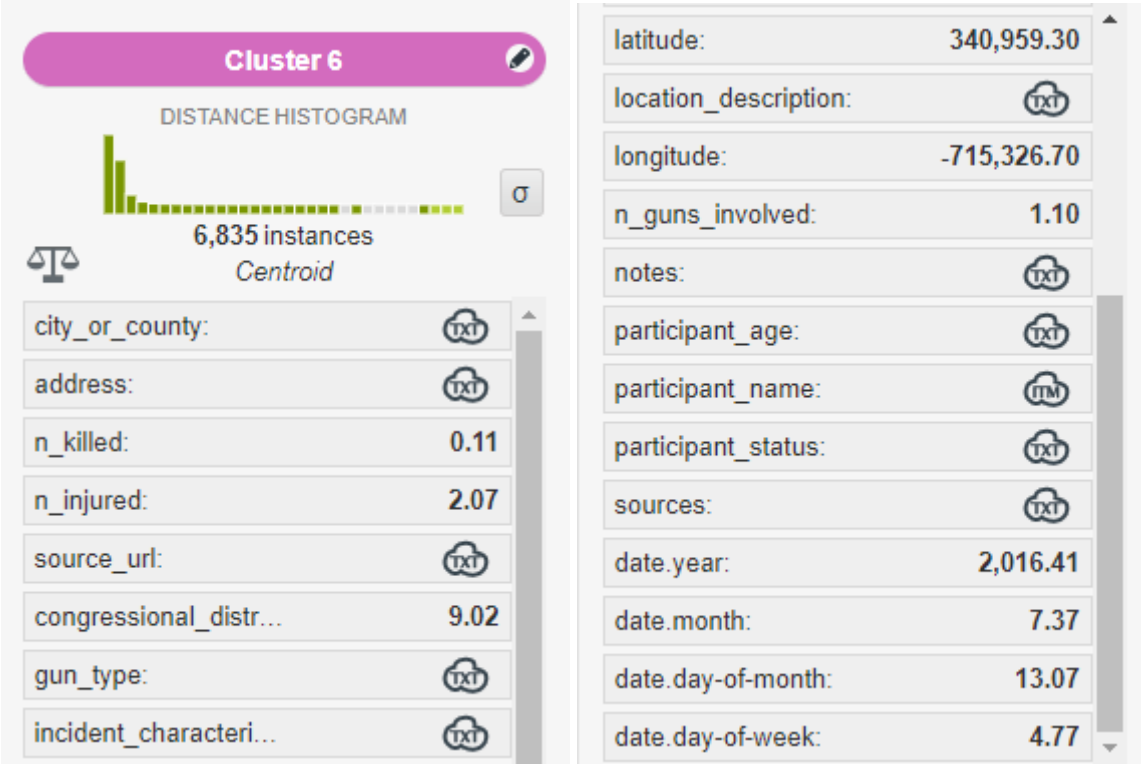
Slika 13: Podaci klastera 1

Klaster 1 je jako sličan klasteru 2, ali ima jednu značajnu razliku, a to je da je u njegovim instancama korišten 2 ili više pištolja iz čega zaključujemo da je postojalo više od jednog napadača, a najvjerojatnije se radi o dva. Moguće je da ste pomislili zašto uopće postoje ta dva klastera kad su skoro pa isti, ali u ovom dijelu ja objašnjavam samo nekoliko atributa, a postoji ih još puno koji utječu na konačni rezultat. Ostale atribute objašnjavati ću u drugim metodama jer će tamo imati puno veću važnost.



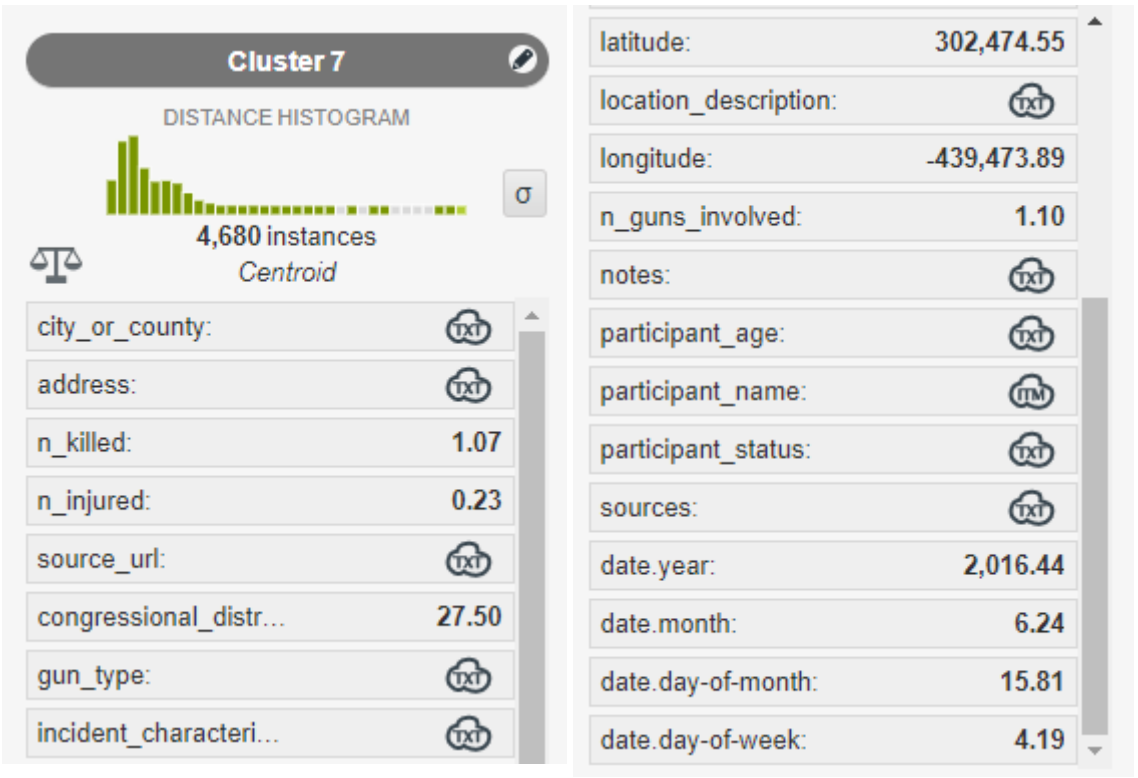
Slika 14: Podaci klastera 0

Klaster 0 je također veoma sličan klasterima 1 i 2 te ga neću previše objašnjavati samo ću naglasiti da pogledamo atribute vezane uz datume. Vidimo da je atribut mjeseca približan svakom od njih te se sve vrti oko sredine godine, ali velika je razlika u danima u mjesecu. Možemo vidjeti da se zločini u klasteru 1 odvijaju početkom, u klasteru 0 sredinom dok u klasteru 2 krajem mjeseca.



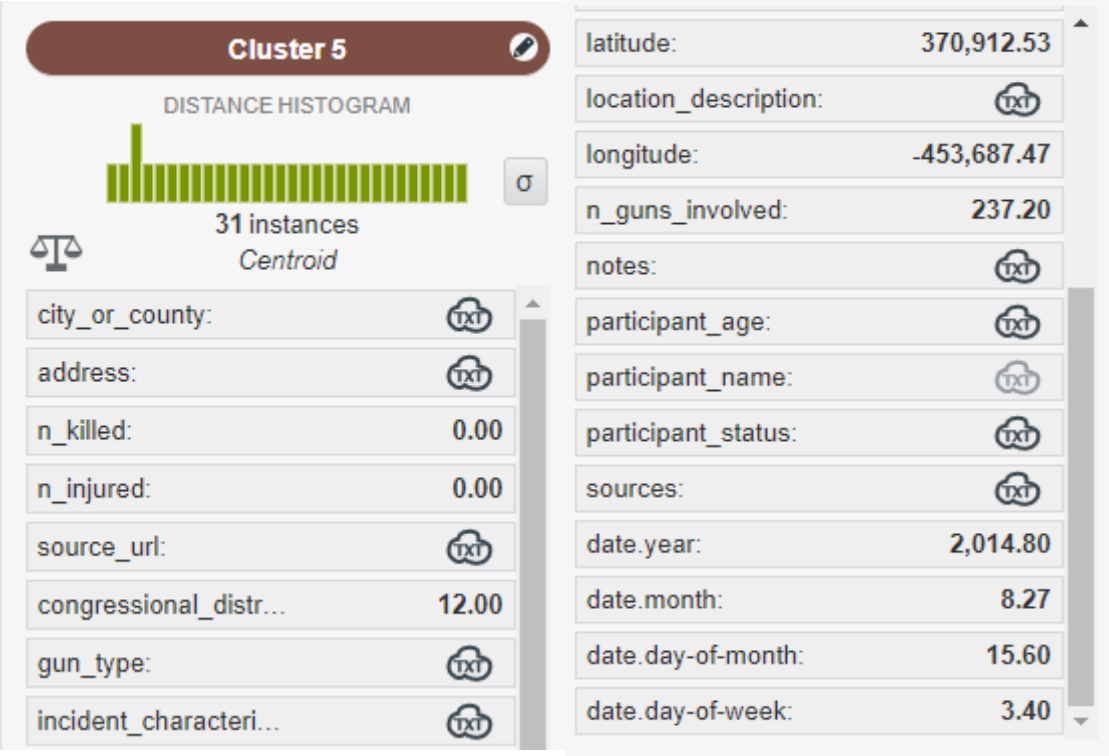
Slika 15: Podaci klastera 6

U klasteru 6 možemo vidjeti da je broj ozlijeđenih veći od dva dok je broj smrtno stradalih malo veći od nule što znači da u ovim zločinima ljudi koji su bili ozlijeđeni su u velikom broju preživjeli.



Slika 16: Podaci klastera 7

Klaster 7 se sastoji od jako malo instanci te nema preveliku važnost na konačan rezultat, ali ipak možemo nešto i iz njega izvući. Vidimo da je broj ubijenih malo veći od jedan dok broj ozlijeđenih iznosi 0.23 te broj uključenih pištolja je 1.10. Možemo zaključiti da je sudjelovao jedan napadač koji je uvijek ubio jednu osobu te moguće da je još jednu ozlijedio.



Slika 17: Podaci klastera 5

Ovo je daleko najmanji klaster u našoj analizi te se sastoji samo od 31 instance. Odmah na prvu možemo vidjeti da ih je alat grupirao po broju uključenih oružja koji je stvarno velik te iznos 237.20, a broj ozlijeđenih i umrlih je točno nula. Iz ovoga možemo zaključiti da se radilo o nekom organiziranom zločinu gdje je sudjelovalo veliki broj ljudi, ali cilj im nije bio ubijanje nego nešto drugo, može biti pljačka, ali može i biti nekakvo slavlje koje organiziraju pojedine kulture pucavši puškama u zrak.

Vjerojatno ste primijetili da u svim osim zadnjem klasteru atribut godine je broj 2016, to ne znači da su se svi ti zločini odvijali u 2016 godini nego klaster analiza uzima prosjek. Pošto je ovo skup podataka rađen nad zločinima između 2013. i 2018. godine, s tim da te dvije godine nisu sudjelovale u potpunosti, jednostavnom matematikom možemo zaključiti da prosjek između tih godina daje 5,5 odnosno 2016. godinu što govori i naša analiza. Isto je tako i za dan u tjednu.

5.5. Stablo odlučivanja na realnim podacima

Stablo odlučivanja predstavlja još jednu od najčešće korištenih data mining metoda analize. Nastalo je na bazi statističkih metoda raspoznavanja uzoraka. Prednost stabla odlučivanja zasigurno je jednostavnost i lako razumijevanje dobivenih rješenja problema za koji smo isto koristili. Stablo možemo koristiti za grupiranje kao i klaster metodu, za predviđanje i procjenu vrijednosti, za razvrstavanje podataka i slične primjene. Standardna i najčešće korištena metoda izrade modela korištenjem stabla je rekurzivno particioniranje. Ono počinje od korijena stabla odnosno tako zvanom top down metodom. Te čvorove stabla nazivamo roditeljima, dok su daljnje podijele čvorovi djeca.

Za naše istraživanje izraditi ćemo nekoliko modela stabla odlučivanja te ćemo u nastavku ovog poglavlja opisati ona najvažnija i iz njih izvući neka pravila te ih objasniti.

5.5.1. Model temeljen na atributu *„n-killed“*

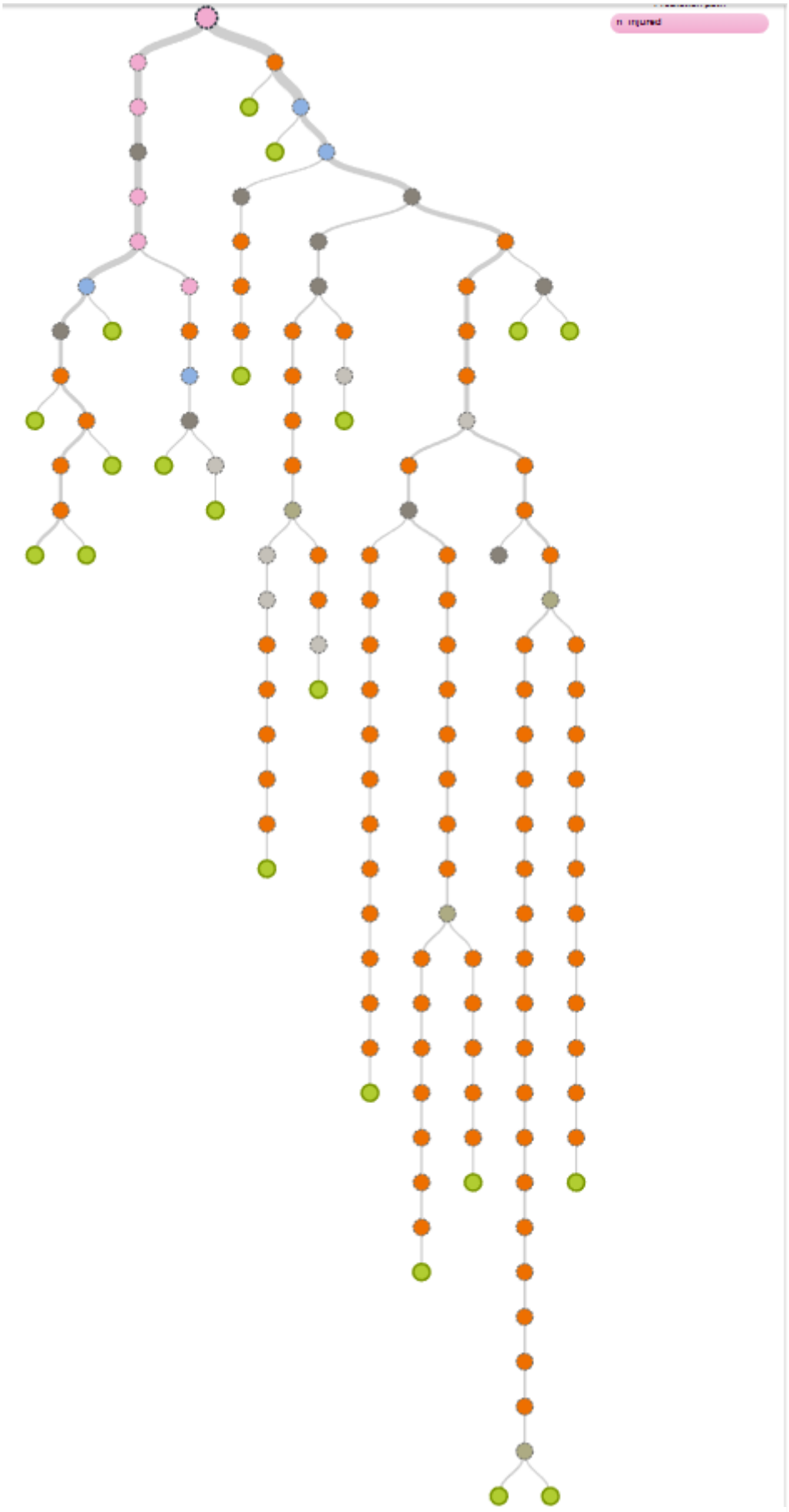
Na slici 17 možemo vidjeti kako cijelo stablo izgleda. Stablo je poprilično veliko, a sastoji se samo od nekoliko atributa koje ću navesti u nastavku te njihovu boju prema slici. Navoditi ćemo samo originalna imena atributa, a detaljnije objašnjeni atribut nalaze se u poglavlju 5.2.2.

Atributi korišteni na modelu su:

- *„n-injured“* – svijetlo roza boja
- *„state“* – narančasta boja
- *„n_guns_involved“* – svijetlo plava boja
- *„date“* – različite nijanse sive boje
- *„n_killed“* – zelena boja

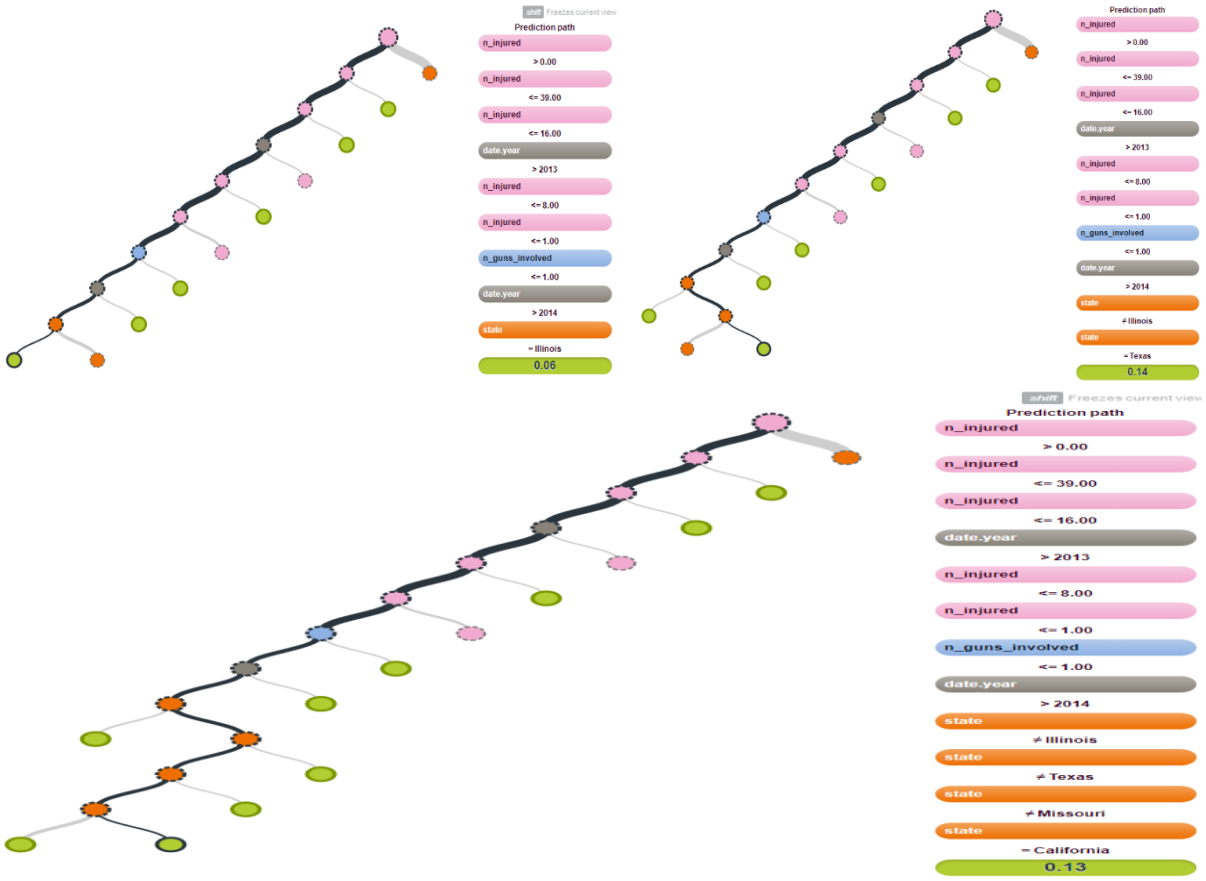
Grananje modela također možemo vidjeti na slici dok pravila su nam prikazana na desnoj strani, ali ih možemo i izvesti u obliku nekog od ponuđenog programskog jezika.

Također, u ovom alatu možemo izračunati i točnosti modela, te točnosti određenih grana, odnosno, pravila, no da bi to mogli moramo imati sve numeričke atribute što na ovom modelu nije sadržano. Stoga točnost modela i pojedinih grana ću prikazati na nekim od sljedećih modela.



Slika 18: Stablo odučavanja na temelju atributa „*n_killed*“

U nastavku ćemo objasniti neka od pravila zajedno sa njihovom slikom.



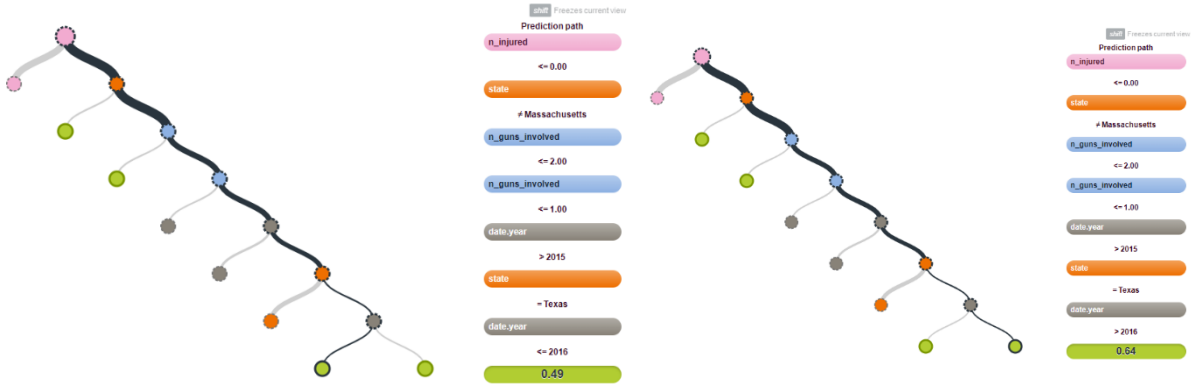
Slika 19: Predikcije o broju ubijenih

Na prethodnoj slici možemo vidjeti zapravo tri slike koje sam grupirao u jednu kako bi što bolje mogao opisati predikcije. Ove slike se razlikuju samo po jednom pravilo, a to je država za koju se radi predikcija. Glavno pravilo ovih grana glasi:

„Ukoliko broj ozlijeđenih manji od jedan, broj uključenih pištolja manji od jedan te se nalazimo u godini većoj od 2014. i u državi…“

Na kraju smo ostavili tri točkice jer se razlikuje pravilo po pojedinoj državi. Uz navedene uvjete u Illinois-u broj ubijenih iznositi će 0.06, u Kaliforniji 0.13 dok u Texasu 0.14. Što nam zapravo govore ove brojke? Recimo da smo mi ministarstvo unutarnjih poslova SAD-a te želimo smanjiti stopu oružanih nasilja u državi. Uz pomoć ovih podataka možemo vidjeti koliko će, uz iste uvjete, biti broj ubijenih po pojedinoj državi. Kako bi smanjili oružana nasilja najbolje nam je pojačati policiju i druge službe u Texasu jer je tamo najveća stopa ubojstava. Osim ovog, ovdje se može izvući još puno odluka i načina kako nešto smanjiti, povećati ili nešto sasvim drugo. Ovdje se radi o vrlo malim brojkama razlike te nas ova pravila mogu

odvesti u skroz krivom smjeru te zbog toga moramo koristiti i ostale modele kako bi nešto sa sigurnošću mogli tvrditi.



Slika 20: Predikcija broja ubojstava u Texasu

Slika 19 nam pokazuje predikciju broja ubojstava u državi Texas, pravilo za obje slike je veoma slično, jedina razlika je u godini događaja. Pravilo za ovu predikciju glasi:

„Ukoliko nema ozlijeđenih, broj uključenih pištolja je manji od 1, nalazimo se u Texasu te je godina…“

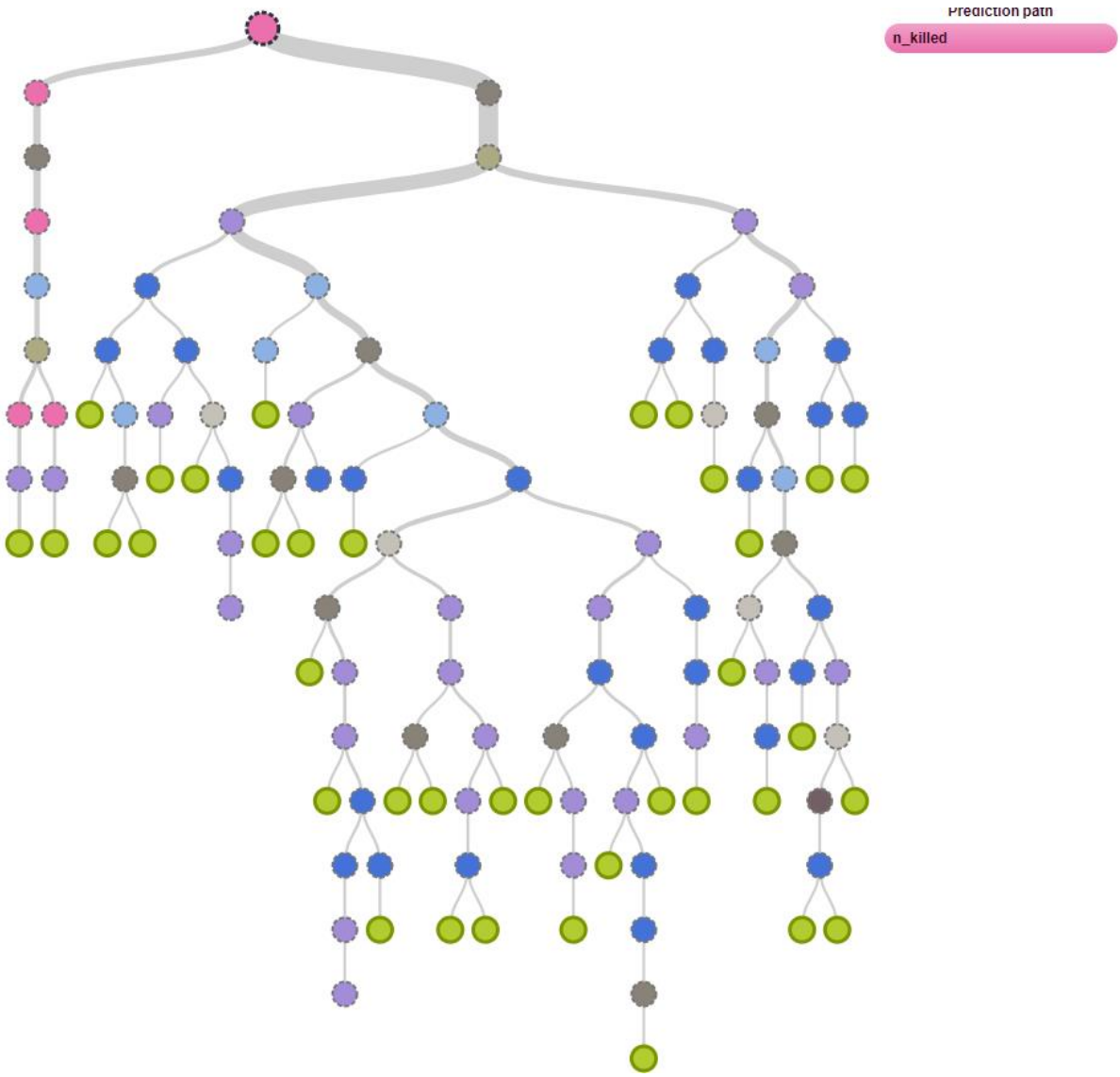
Opet smo tri točkice iskoristili kao nastavak pravila u kojemu mijenjamo samo prije i poslije 2016. godine. Ukoliko je godina manja od 2016. broj ubijenih iznosi 0.49 dok za godinu veću od 2016 iznosi 0.64. Iz ovoga da se zaključiti da broj ubijenih je porastao za puno u odnosu na jednu godinu te možemo očekivati trend porasti ubojstava u Texasu što nikako nije dobro. Ministarstvo unutarnjih poslova bi svakako morala iskoristiti ovaj podataka te snažno utjecati na smanjenje broja ubojstava u budućnosti svojim metodama koje oni najbolje znaju, ali neke od njih bi bile povećanje policijskih djelatnika na tom području, povećavanjem kazni za pokušaj ili počinjeno ubojstvo, itd.

5.5.2. Model temeljen na atributu „*n-injured*“

U ovom modelu smo koristili samo numeričke atribute kako bi mogao napraviti izračun točnosti modela. Na slici 20 možemo vidjeti kako izgleda ovo cijelo stablo, a u nastavku ću objasniti od kojih atributa se sastoji.

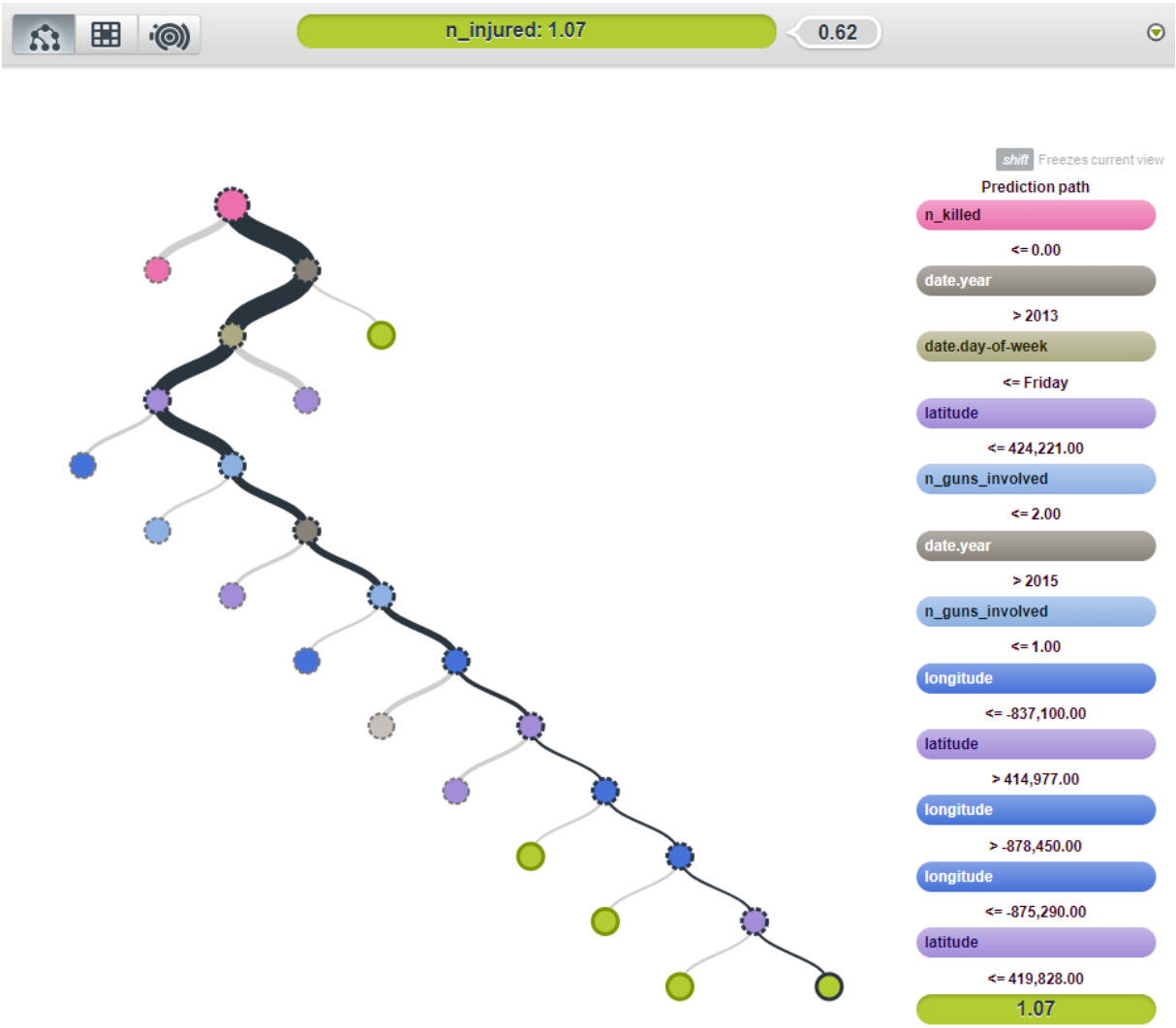
Korišteni atributi su:

- *„n-killed“* – svijetlo roza boja
- *„latitude“* – svijetlo ljubičasta boja
- *„longitude“* – plava boja
- *„date“* – različite nijanse sive boje
- *„n_guns_involved“* – svijetlo plava boja
- *„n_injured“* – zelena boja



Slika 21: Stablo odlučivanja na temelju atributa "*n_injured*"

Nastaviti ćemo ovo poglavlje kao i prethodno sa slikama nekih od grana te objašnjavanja njegovih pravila i donijeti zaključke temeljem toga.



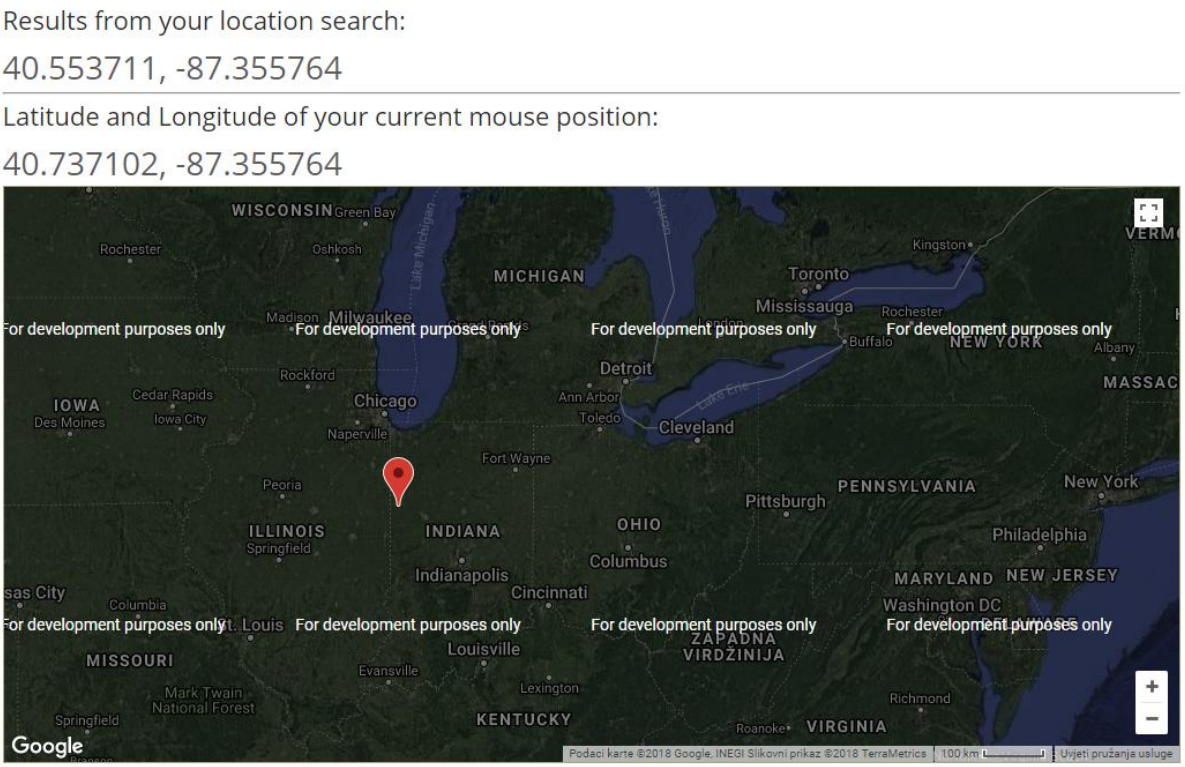
Slika 22: Predikcija broja ozlijeđenih

Prethodna slika nam pokazuje pravilo koje ću u nastavku objasniti, ali htjeli bi prvo spomenuti da gore pored atributa broja ozlijeđenih možemo vidjeti točnost ovog pravila, za ovo pravilo ono iznosi 0.62, odnosno, 62% da je točno. Pravilo za grananje prikazano na slici glasi:

„Ako nemamo ozlijeđenih te smo u godini većoj od 2015., dan u tjednu je ponedjeljak-petak, nalazimo se između 419,828 i 424,221 zemljopisne širine i između -878,450 i -875,260 zemljopisne dužine te je broj uključenih oružja je manji od 1, broj ozlijeđenih iznositi će 1,07“

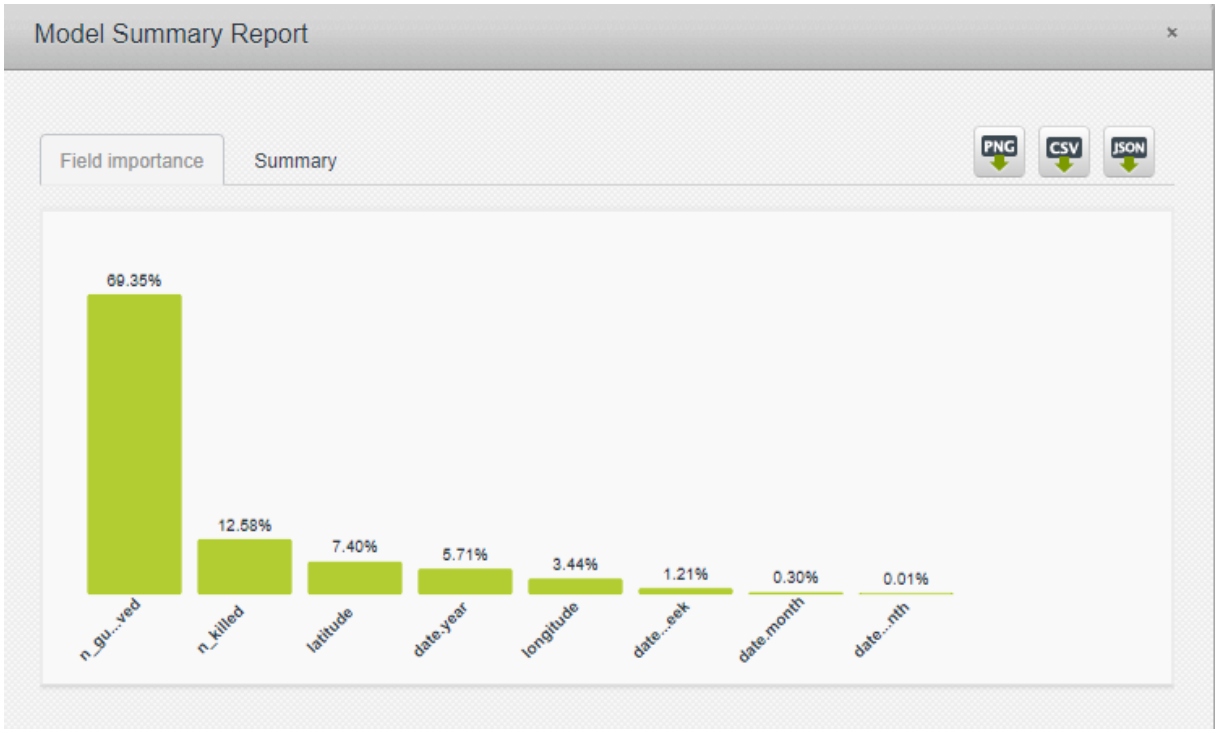
U ovom modelu smo umjesto kategorijskog atributa država koristili zemljopisnu širinu i dužinu te ukoliko iznesene vrijednosti unesemo u kartu možemo dobiti točno o kojoj lokaciji se radi. Brojeve zemljopisne visine i širine treba podijeliti sa brojem tisuću da bi dobili stvarne koordinate te lokaciju za koje vrijedi ovo naše pravio. Kada sve to uzmemo u obzir možemo vidjeti lokaciju na slici 22 koja prikazuje da je to lokacija između država Indiane i Illinoisa gdje

se u blizini nalazi i Chicago koji je imao najveći broj ubojstava i u prethodnom modelu. Ovi podaci i pravila bi također koristilo ministarstvu unutarnjih poslova koji bi mogli smanjiti broj ozlijeđenih na navedenom području jer vidimo da broj ozlijeđenih je također veći na područjima gdje je i veća stopa ubojstava te možemo sa sigurnošću reći da su to mjesta većeg kriminala i većeg broja zločina.



Slika 23: Prikaz lokacije uz pomoć koordinata

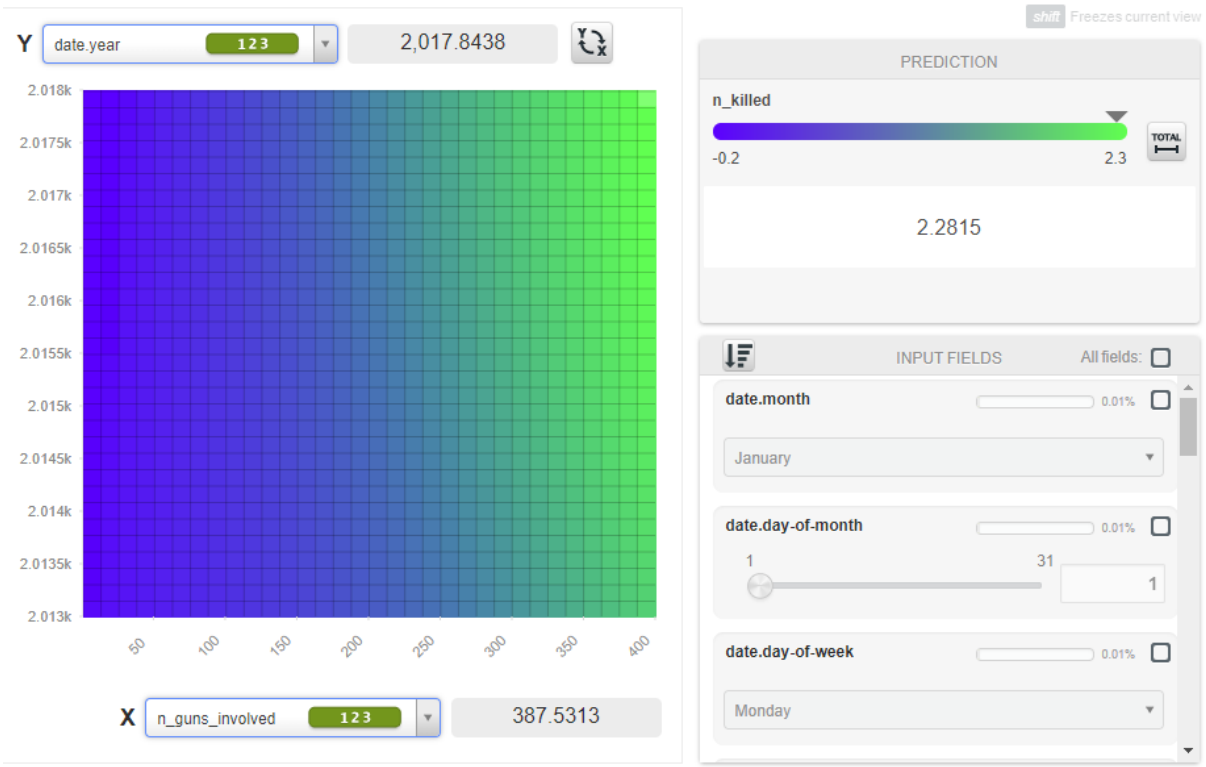
Na slici ispod možemo vidjeti koji atributi su imali najviše utjecaja na ishod pravila. Iz našeg izvješća vidimo da atribut *„n_guns_involved“* ima daleko najveći utjecaj što je i logično jer da nema oružja onda se ne bi ni smatralo oružanim nasiljem. Ostali atributi imaju slične vrijednosti utjecaja s time da *„n_killed“* predvodi dok *„date“* se nalazi daleko na zadnjem mjestu.



Slika 24: Utjecaj atributa na model stabla odlučivanja

5.6. Neuronske mreže na realnim podacima

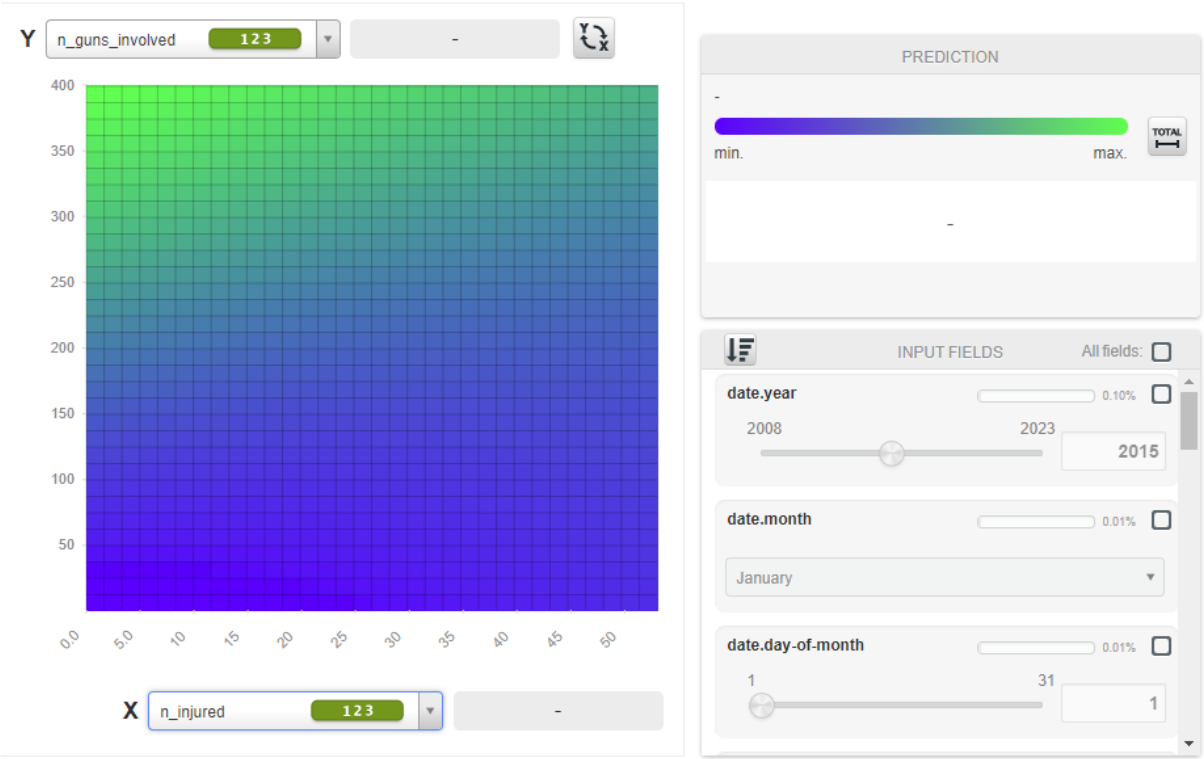
Neuronske mreže predstavljaju računalne strukture koje se temelje na realnom svijetu i funkcioniraju slično kao neuroni kod ljudi. Za izrade neuronske mreže uzeli smo atribut *„n_killed“* jer smatramo da je on od najveće vrijednosti za donošenje odluka. U primjeru koji obrađujemo prikazivati će se samo dvije boje, a to su plava i zelena. Plava boja će prikazivati mali broj ubojstava dok će zelena prikazivati veliki broj. Važno je još napomenuti da neuronske mreže prihvaćaju samo numeričke vrijednosti te da ću u nastavku ponovno koristiti zemljopisnu visinu i širinu umjesto naziva država.



Slika 25: Neuronska mreža od atributa "*n_guns_involved*" i "*date.year*"

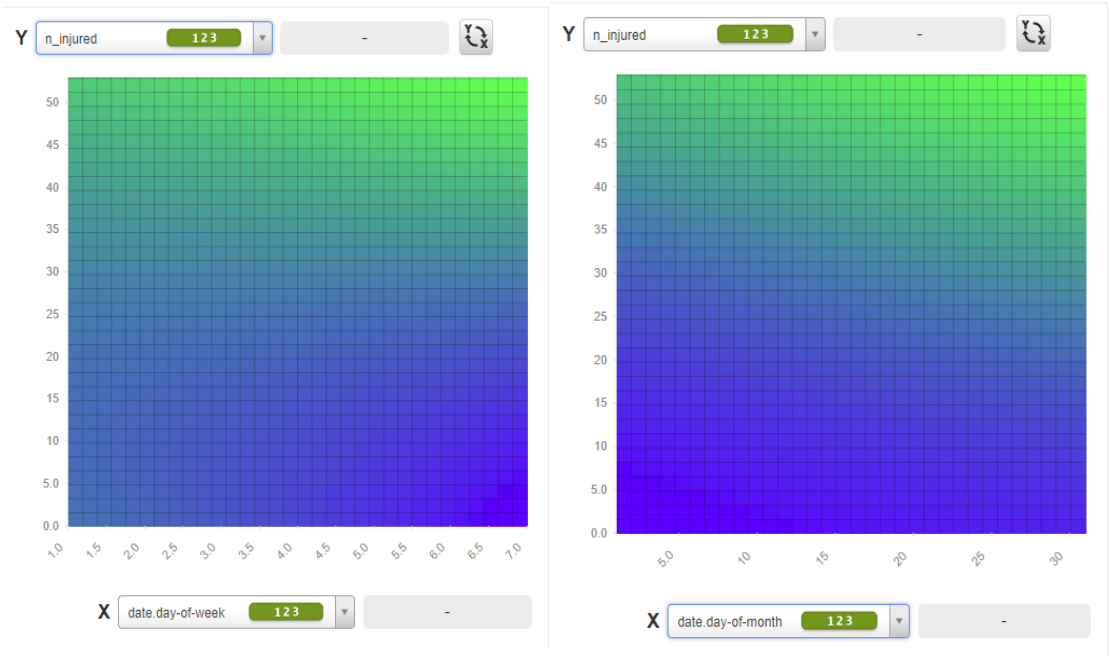
Slika iznad nam pokazuje predikciju broja ubijenih na temelju godine i broja korištenih pištolja u zločinu. Vidimo da su nam s desne strane sve kućice zelene boje što označava veliki broj ubojstava, a razlog tome je povećanje broja korištenih oružja što je i sasvim logično. Međutim kada zadnjim stupcem krenemo prema gore možemo vidjeti da dobivamo svijetlu nijansu zelene boja što nam označava povećanje ubojstava. Iz ove dvije činjenice možemo zaključiti da broj uključenih oružja u zločinu ima veliki utjecaj na broj ubojstava, ali da s godinama također taj broj raste no ipak ne u toliko mjeri kao kod prvog atributa. Vjerojatno ste čuli da Američko ministarstvo unutarnjih poslova ima kampanju skupljanja oružja u policijskim

postaja gdje Vas apsolutno ništa ne pitaju prilikom predaje oružja nego je samo bitno da oružje nije više u Vašim rukama. Iako mnogi smatraju tu kampanju glupom i beskorisnom ipak iz ovih podataka možemo vidjeti da to i nije baš beskorisno jer smanjenje oružja moglo bi utjecati i na smanjenje ubojstava.



Slika 26: Neuronska mreža od atributa "*n_guns_involved*" i *"n_injured*"

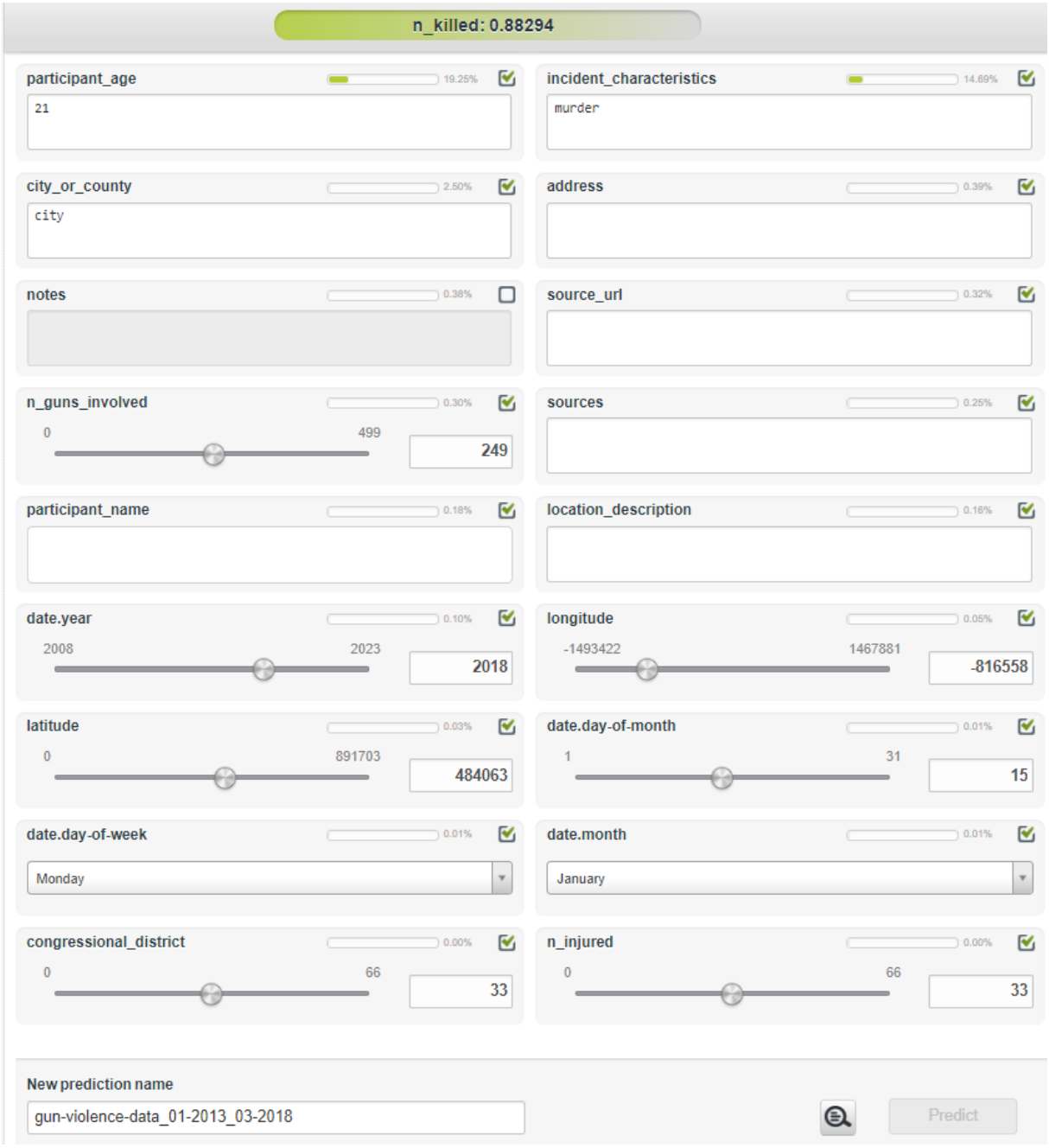
Gore imamo veoma zanimljiv graf koje prikazuje odnos između broja korištenih oružja i broja ozlijeđenih. Kažem da je zanimljiv iz razloga što možemo vidjeti kako broj ozlijeđenih raste tako broj ubijenih pada. Iz te činjenice dajemo si za pravo zaključiti da je u slučaju nasiljem oružjem vjerojatno bilo pokušaja ubojstva, ali neki su ipak uspjeli preživjeti. Drugim riječima, napadači su zapucali iz oružja na mete, ali su ih samo ozlijedili što dovodi do toga ukoliko imamo ozlijeđenu osobu u slučaju mala je vjerojatnost da imamo i ubijenu dok ako nemamo ozlijeđenih velika je šansa da je netko ubijen.



Slika 27: Neuronska mreža nad atributima "*date*"

Na slici 26 gledamo samo x-os, odnosno, usporednu atributa *„date.day-of-week“* i *„date.day-of-month“*. Nad danom u mjesecu nema prevelike razlike između broja ubojstava, ali ipak možemo vidjeti da na kraju mjeseca ipak taj broja malo poraste. Nad danom u tjednu ipak je veća razlika kako idemo prema kraju tjedna. Iz toga možemo naslutiti da je stopa ubojstava veća za vrijeme vikenda te za to može biti mnoštvo razloga. Neki od njih mogu biti da se ljudi vikendima okupljaju na većim i popularnijim mjestima te ukoliko dođe do nekog velikog napada vjerojatnost da više osoba bude ubijeno je veća nego radnim danima. Također, za dan u mjesecu opet može biti puno razloga te bi trebali provoditi još daljnje analize kako bi za sigurnošću mogli izvući neka znanja te uz pomoć njih probati smanjiti tu stopu ubojstava i općenito količinu oružanog nasilja.

Uz pomoć neuronskih mreža možemo odrediti točnu predikciju na temelju rezultata. Za našu predikciju unijeli smo podatke o sebi te gdje će se zločin dogoditi, da će opis događaja biti „ubojstvo“ i da godina događaja biti će trenutna(2018). Rezultat predikcije nam je izbacio da broj ubijenih na događaju pod takvim uvjetima je 0.88 što je poprilično velik broj, ali moramo uzeti u obzir da smo napisao izričito ubojstvo u opis događaja. Rezultat možete pogledati na slici 27.



Slika 28: Predikcija broja ubojstva nad mojim podacima

Ovime smo htjeli pokazati da su neuronske mreže veoma moćne te da uz pomoć njih može puno toga pročitati što nam inače nije vidljivo ljudskim okom. Ministarstvo unutarnjih poslova bilo koje države, ne samo SAD-a, može i mislim da bi trebala koristiti se prethodim slučajevima te skupljati što više podataka vezanih uz njih kako bi mogla u budućnosti predvidjeti kada i gdje bi se mogao dogoditi napad te uz pomoć toga donijeti odluke kako ga spriječiti, odnosno, kako svesti posljedice na minimum. Ovo je samo jedan od mnogobrojnih primjera gdje bi se mogle koristi neuronske mreže.

6. Zaključak

U ovom radu pokazali smo te objasnili dva pojma koji imaju svijetlu budućnost te ni jedna veća organizacija neće moći poslovati bez njih. Prvi pojam se odnosi na rudarenje podataka koje se pokazalo kao iznimno dobra metoda za analiziranje velikih količina podataka. Kako bi analiza bilo što kvalitetnija potrebni su prije svega kvalitetni podaci, ali i upotreba odgovarajuće metode je od iznimne važnosti. Danas postoje puno različitih metoda i alata za obradu, analizu i prikaz podataka. Metode koje su zastupljene u ovom radu su samo neke po meni najvažnije od puno preostali koje također imaju svoje prednosti i mane. Rudarenje podataka se pokazalo toliko korisnim da se danas koristi u svakom mogućem područje te se očekuje još dodatni porast njegovom korištenja tijekom budućnosti. Stručnjaci zaduženi za to područje su veoma traženi na tržištu rada, a oni uz pomoć rudarenja podataka izrađuju što optimalnija rješenja na temelju kojih rješavaju nekakve dvojbe ili unaprjeđuju postojeći sustav. Drugi pojam koji smo obradili u ovom radu je upravljanje znanjem koje je složeni postupak te nije primjenjivo niti potrebno u svim organizacijama. Međutim, one organizacije koje ga koriste ono im stvara kompetitivnu prednost, ali moraju ulagati velike napore u njegov razvoj. Drugim riječima, upravljanje znanjem je niz međusobnih povezanih aktivnosti koje mu je svrha maksimizirati efektivnost organizacijskih aktivnosti vezanih uz znanje.

Spojem ova dva pojma dobivamo veoma jak *„alat“* kojim možemo doći do kvalitetnih odluka i zaključaka koji sigurno ne bi bili mogući bez njih. Osim raznih veliki organizacije koje koriste ova dva pojma, postoji još puno načina i objekata koji bi hi mogli koristi. U ovome radu pokazao sam način kako bi ih Ministarstvo unutarnjih poslova SAD-a trebalo koristiti te što sve može postići uz pomoć njih.

7. Literatura

Knjige:

[1] Becerra-Fernandez I, Gonzalez A, Sabherwal R (2004). *„Knowledge Management“*. New Jersey: Prentice Hall

[2] Woods, J.A., i Cortada, J. (2000). „*The Knowledge Managment Yearbook"*

[3] Groff R, Jones P, (2003), „*Introduction to Knowledge Management: Knowledge Management in Business“*. Burlington: Butterworth-Heinemann

[4] Bahtijarević-Šiber F, Sikavica P(2001). „*Leksikon menedžmenta“*. Zagreb: Masmedia

[5] Žugaj M, Schatten M (2005). „*Arhitektura suvremenih organizacija“.* Varaždin: Varaždinske Toplice: Tonimir, Fakultet organizacije i informatike.

[6] Wiig K. M. (2004). „*People-focused knowledge management: how effective decision making leads to corporate success“*. Oxford: Elsevier Butterworth-Heinemann

[7] Snowden D. (2003). „*Innovation as an objective of knowledge management. Part I: Thelandscape of management, Knowledge Management Research & Practice“*. Prosinac 2003, Volume 1, Number 2, Pages 113-119. Palgrave Macmilan Journals

[8] Garača Ž, Jadrić M (2011). „*Rudarenje podataka: različiti aspekti informacijskog društva“*. Split: Ekonomski fakultet u Splitu

[9] Berry M, Linoff G (2004). „*Data mining tehniques, For marketing, sales, and customer relationship managment“*. Indiana: Wiley Publishing Inc

[10] Han J, Kamber M (2006). „*Data mining, Concepts and Techniques“*. San Francisco: Elsevier Inc.

[11] Zekić-Sušac M, Frajman-Jakšić A, Drvenkar N. (2009). „*Neuronske mreže i stabla odlučivanja za predviđanje uspješnosti studiranja“* Ekonomski vjesnik, 22 (2).

Internetske stranice:

[1] Srića V. (2018). „*Otkrivanje znanja iz podataka“*. Preuzeto 25. srpnja 2018. s https://www.scribd.com/document/378101618/05-Otkrivanje-Znanja-Iz-Podataka

[2] Pejić-Bach M. (2005). „*Rudarenje podataka u bankarstvu“*. Preuzeto 25. srpnja 2018. s http://hrcak.srce.hr/file/41477

[3] Klepac G. (2006). „*Što je to data mining ?*“. Preuzeto 25. srpnja 2018. s http://www.goranklepac.com/index.asp?j=HR&iz=1&sa=1&vi=1&hi=1

[4] Gamberger D, Marić I, Šmuc T. (2018). „*Otkrivanje znanja i obrada podataka*“. Preuzeto 26. srpnja 2018. s http://dms.irb.hr/tutorial/hr_tut_dtrees.php

[5] Gamberger D., Šmuc T. (2001). „*Poslužitelj za analizu podataka*“. Zagreb, Hrvatska: Institut Ruđer Bošković, Laboratorij za informacijske sustave. Preuzeto 25. srpnja 2018. s http://dms.irb.hr/

[6] Ujević F. (2004). „*Postupci analize podataka u izgradnji profila korisnika usluga*“. Magistarski rad. Sveučilište u Zagrebu, Fakultet elektrotehnike i računarstva. Preuzeto 25. srpnja 2018. s http://www.maturskiradovi.net/forum/attachment.php?aid=1442

[7] Cindrić M. (2016). *„Rudarenje podataka za društvene analize“.* Završni rad. Sveučilište u Zagrebu, Fakultet organizacije i informatike. Preuzeto 27. srpnja sa https://repozitorij.foi.unizg.hr/islandora/object/foi:1372/preview

[8] Krišto I. (2013). *„Primjena rudarenja podataka u sigurnosti IS-a“.* Završni rad. Sveučilište u Zagrebu, Fakultet organizacije i informatike. Preuzeto 27. srpnja sa https://repozitorij.foi.unizg.hr/islandora/object/foi:652/preview

[9] Novak K. (2014). *„Upravljanje znanjem kao temeljnim resursom u suvremenom poslovanju“.* Diplomski rad. Sveučilište u Zagrebu, Fakultet organizacije i informatike. Preuzeto 27. srpnja sa https://repozitorij.foi.unizg.hr/islandora/object/foi:753/preview

[10] Blažinić V. (2015). *„Utjecaj upravljanja znanjem na zaposlenike“.* Završni rad. Sveučilište u Zagrebu, Fakultet organizacije i informatike. Preuzeto 27. srpnja sa https://repozitorij.foi.unizg.hr/islandora/object/foi%3A1506/datastream/PDF/view

[11] Lepan J. (2012). *„Upravljanje znanjem u suvremenim organizacijama“.* Završni rad. Sveučilište u Zagrebu, Fakultet organizacije i informatike. Preuzeto 27. srpnja sa https://repozitorij.foi.unizg.hr/islandora/object/foi:2247/preview

[12] Policki M. (2009). *„Proces data mininga nad podacima o prodaji tekstila“.* Završni rad. Sveučilište u Zagrebu, Fakultet organizacije i informatike. Preuzeto 27. srpnja sa https://repozitorij.foi.unizg.hr/islandora/object/foi:2024/preview

[13] Đurić Okreša B. (2017). Prezentacije iz kolegija *„Upravljanje znanjem“*. Sveučilište u Zagrebu, Fakultet organizacije i informatike. Preuzeto 27. srpnja sa portala za e-učenje *„Moodle.com“.*

Alati:

[1] BigML, Inc. (2018). *„About BigML“* . Preuzeto 2. kolovoza sa https://bigml.com/about

[2] Kaggle Inc. (2018). *„Gun violence data“*. Preuzeto 2. kolovoza sa https://www.kaggle.com/jameslko/gun-violence-data

8. Popis tablica i slika

8.1. Popis slika

8.2. Popis tablica